2026
State of Application Strategy Report

CONTENTS

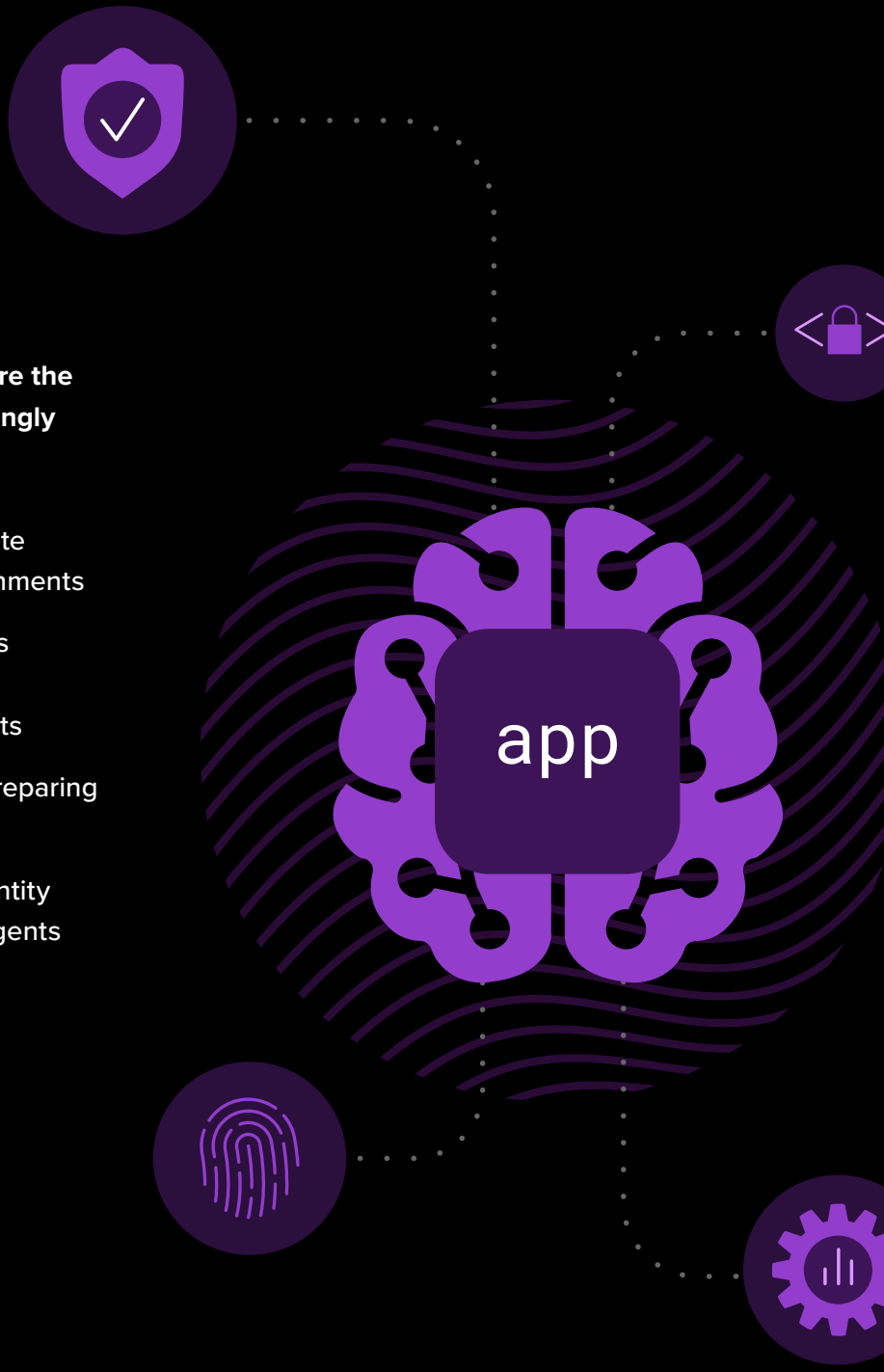
KEY 2026 FINDINGS

Inference is becoming a distributed system in the AI application journey

- Organizations use an average of seven AI models in production
- 78% of organizations are operating AI inference themselves
- 52% are using multi-model orchestration to adapt and extend their AI models
- 64% allow AI to adjust policies or configurations
- Application services for prompt handling have the single greatest operational impact

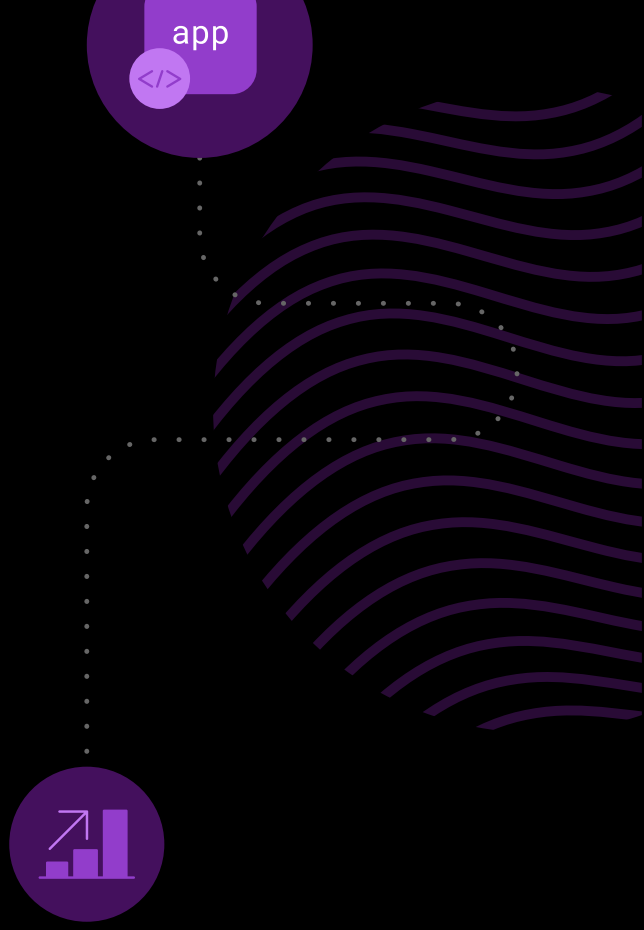
Hybrid multicloud strategies are the operating reality in an increasingly AI-powered threat landscape

- 93% of organizations operate in hybrid multicloud environments
- 86% run applications across on-premises, public cloud, and colocation environments
- 98% of organizations are preparing for agentic AI
- 77% expect issues with identity and access control for AI agents



INTRODUCTION

Welcome to the inference age



With organizations using an average of seven AI models, AI inference has crossed the type of threshold that enterprises tend to recognize only in hindsight.

The results of the 2026 F5 State of Application Strategy Survey demonstrate how activities that began as AI experimentation only a few years ago have quietly become part of organizational architectures. This shift is not defined by the sophistication of AI models, but by how inference is now embedded in production systems, operational decision loops, and the pipelines for delivering apps, products, and services.

More than three-quarters of respondents are operating their own inference, and a similar 77% report inference as their dominant AI activity versus model training or tuning. The AI economy has moved from building to operating, and those organizations are no longer simply “using” AI. They are managing their businesses through it, shifting it from innovation to infrastructure. Furthermore, the data shows they are doing so at scale despite significant challenges like those reported by 88% of respondents.

One indicator of this transition is that survey respondents report simultaneously operating or actively evaluating an average of seven AI models. This proliferation in the use of models is driven not by novelty or by a failure of any one model to do the job. Rather, businesses are turning to a variety of models to meet needs related to API compatibility, costs, availability, data sovereignty, and compliance requirements.

In short, we’ve entered a multi-model world, where inference stops behaving like a single endpoint and starts behaving like a distributed system.

You thought multicloud application delivery was complex? Meet multi-model AI inferencing. AI inferencing across multiple models raises the same architectural and security challenges as any distributed production workload. The primary hurdle is no longer model capability but organizational and architectural scalability.

78%

operate their own inference service

Enterprises operate an average of

seven AI models

Scalability becomes an issue because distributed inferencing shifts the center of gravity away from the model itself and toward the systems that govern how inference traffic is shaped, routed, constrained, and observed. As with distributed app deployment, multiple tools may be used to manage those systems, adding complexity. Meanwhile, new domains of responsibility —such as model-aware authentication, semantic abuse detection, token-based cost governance, and others—arise.

But these new inference responsibilities do not align neatly with the existing remits of platform teams, security teams, or application teams. The new inference domains cross them. Nor is the typical enterprise response to quietly ask existing teams to absorb the new and unfamiliar work. Rather, enterprises differentiate, creating new teams with their own preferred tools. Without deliberate consolidation or convergence, the resulting complexity and security risks become fractal. The challenges repeat across delivery, security, reliability, and governance layers.

Our data already shows consolidation lagging: only 28% of organizations report streamlining developer workflows through a single AI management point. The rest manage across fragmented control surfaces even as inference becomes a business-critical operation.

Nonetheless, organizations are rapidly moving forward in adapting multi-model, distributed inferencing to optimize how AI empowers their teams. How they manage the resulting complexity could significantly shape the outcomes of their AI deployments now and into the future.

These findings together lead to a single conclusion: If AI fails, stalls, or becomes prohibitively expensive, it will not be because models are insufficient. Failure will come to organizations that underestimate the operational gravity, complexity, and security risks that AI introduces. Unless decision makers get ahead of point solutions, default services that apply only to parts of their AI portfolios, and traditional solutions to AI-era problems, AI complexity will outpace their ability to operate and secure it.

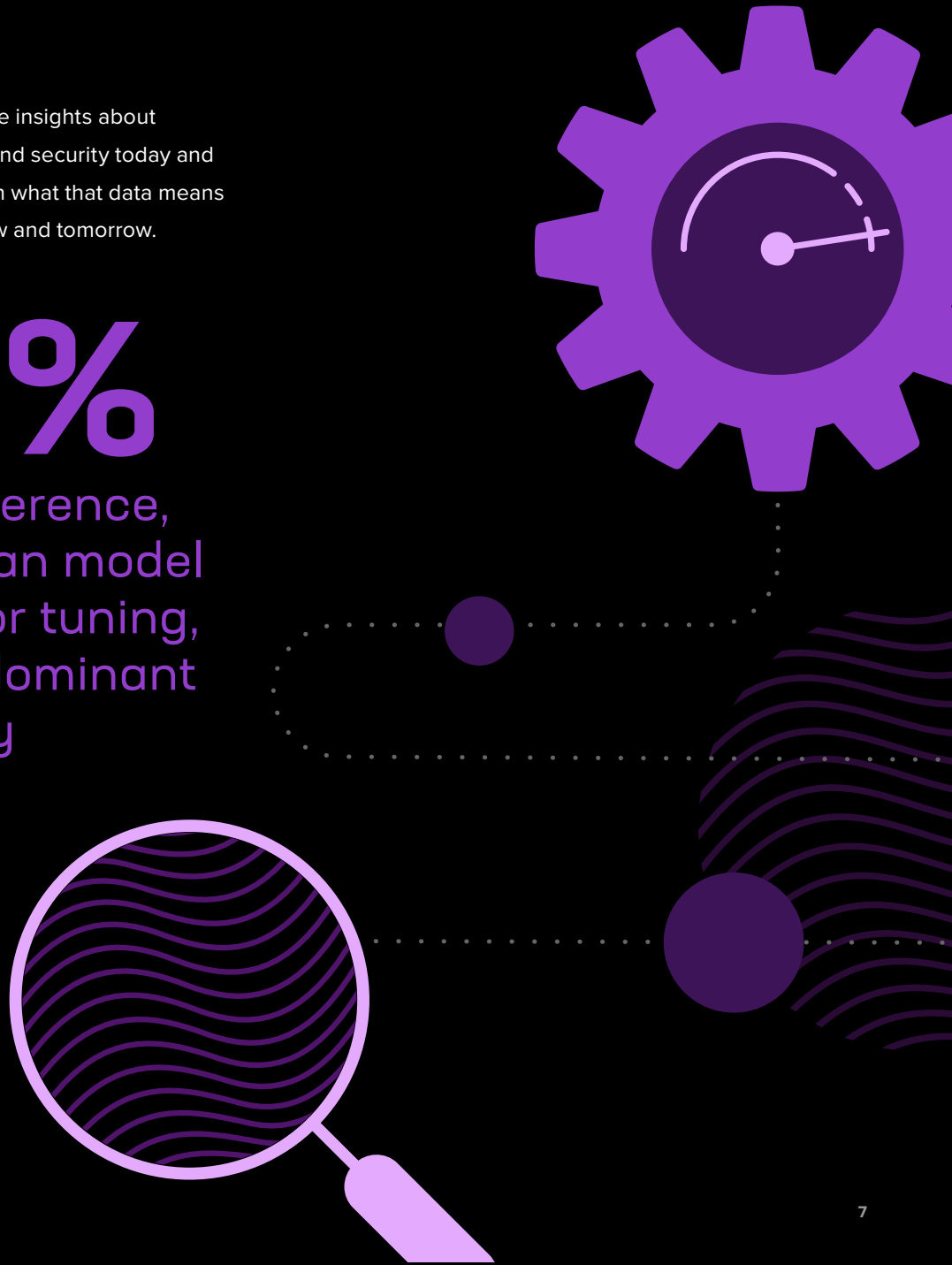
Complexity and security risks become fractal

On the other hand, by intentionally converging controls and empowering teams with cross-model observability and protection, organizations can reap the same efficiency, performance, and innovation advantages from multi-model AI as they are already achieving with hybrid multicloud application distribution.

Keep reading for more insights about application delivery and security today and the F5 perspective on what that data means for digital leaders now and tomorrow.

77%

report inference, rather than model training or tuning, as their dominant AI activity



SECTION 1

Inference has arrived

Modernization has turned to AI

In some ways, the landscape of digital transformation has stabilized.

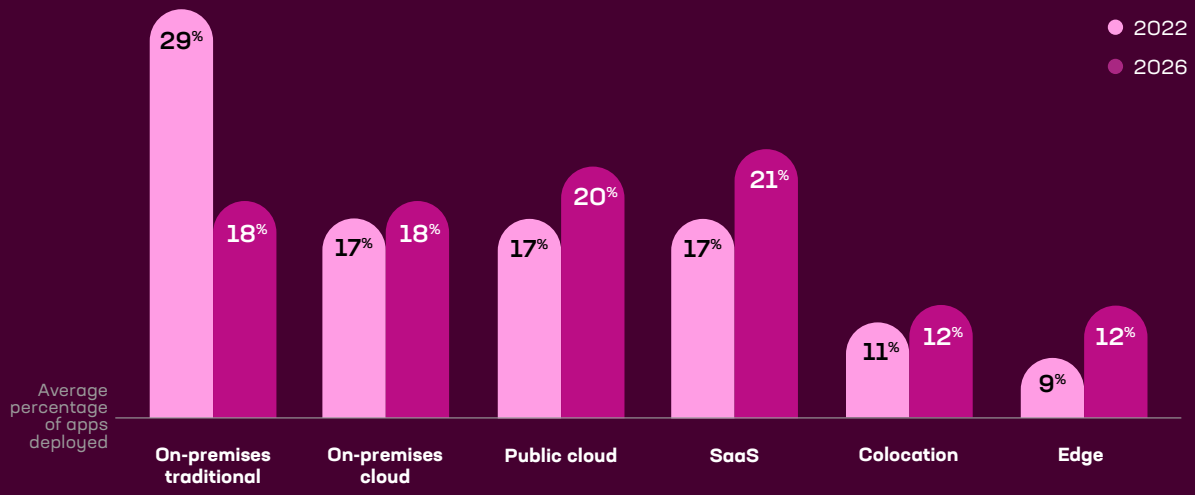
Organizations have modernized traditional apps into on-premises cloud apps or Software as a Service (SaaS), changing how those applications are operated to reduce friction. The push to modernize has slowed, although not stalled. In 2020, modern apps represented 29% of the average application portfolio. Today that figure is 53%, the same as in 2025, in part because over the last year, organizations have largely turned their attention to AI. We expect the percentage of modern apps in the average portfolio to continue to climb, if at a somewhat slower rate. In 2023, 16% of respondents told us they had no plans to retire their remaining traditional apps. Three years later, that floor has nearly been reached.

86% run applications across on-premises, public cloud, and colocation environments

APP DISTRIBUTIONS HAVE LEVELED

We asked:
Of all your applications deployed, roughly what percentage are utilizing the following locations/ deployment models?

We learned:
Apps have settled almost uniformly across hybrid deployment environments and models, and few of the remaining traditional apps will be retired.



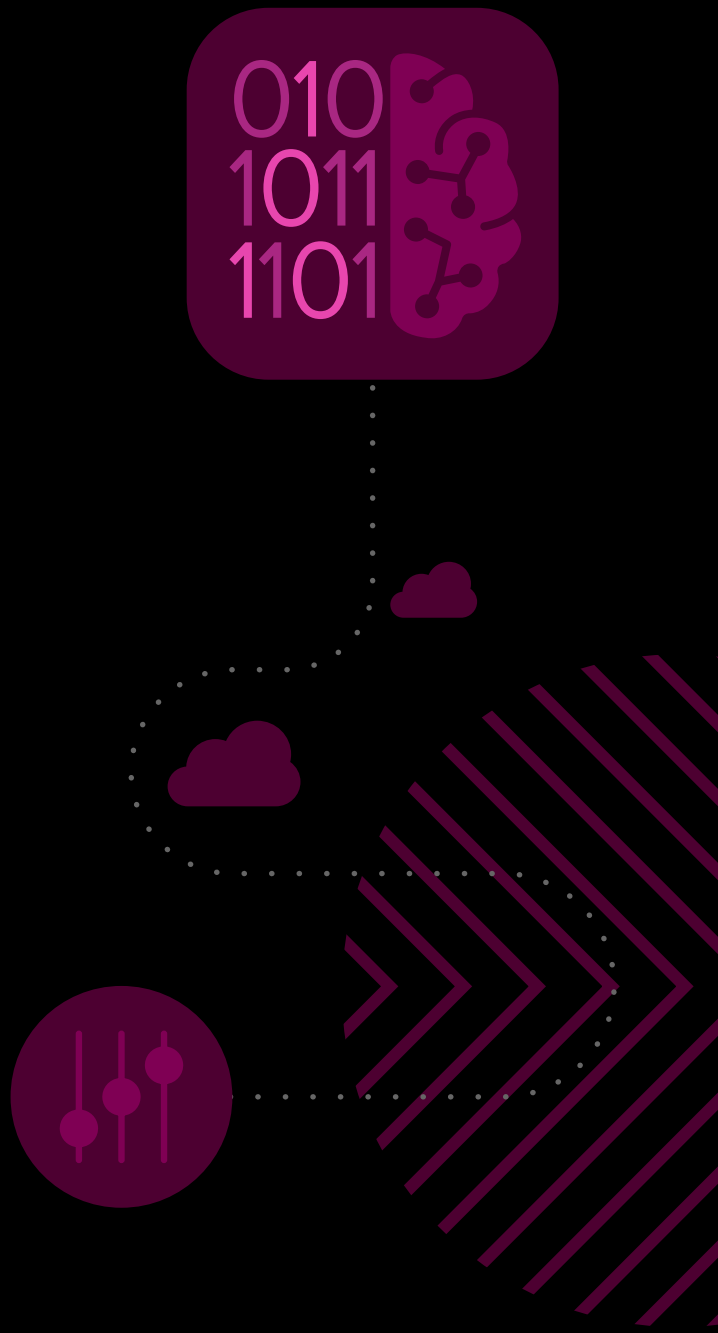
More than nine in 10 organizations (93%) manage hybrid multicloud environments, and while app migration has slowed somewhat, the diversity of hosting locations continues to explode. Respondents to our survey manage five of their own data centers, on average, plus an additional five colocation facilities for a total of 10 infrastructure locations—not to mention four different public cloud providers used by the average organization.

Hybrid multicloud deployments are here to stay

Amid this complexity, a great deal of action surrounds AI. The years of prototypes, proofs of concept, and cautious experimentation are over. Instead, organizations are running AI as an integral part of their businesses. They're doing so at scale despite reporting significant challenges. Inference has moved from novelty to necessity, and more than three-quarters of survey respondents are operating their own inference service. Most are managing more than one. The average number of inferencing services is two, and one in five survey respondents manages three or more.
Most organizations want to control and customize at least some of their services. Although 59% of respondents use public instances from cloud service providers, nearly as many use private instances or manage their own open-source instances. Only 36% rely primarily on public AI as a Service. Only 3% use no AI inferencing service at all.

93%

manage hybrid multicloud environments



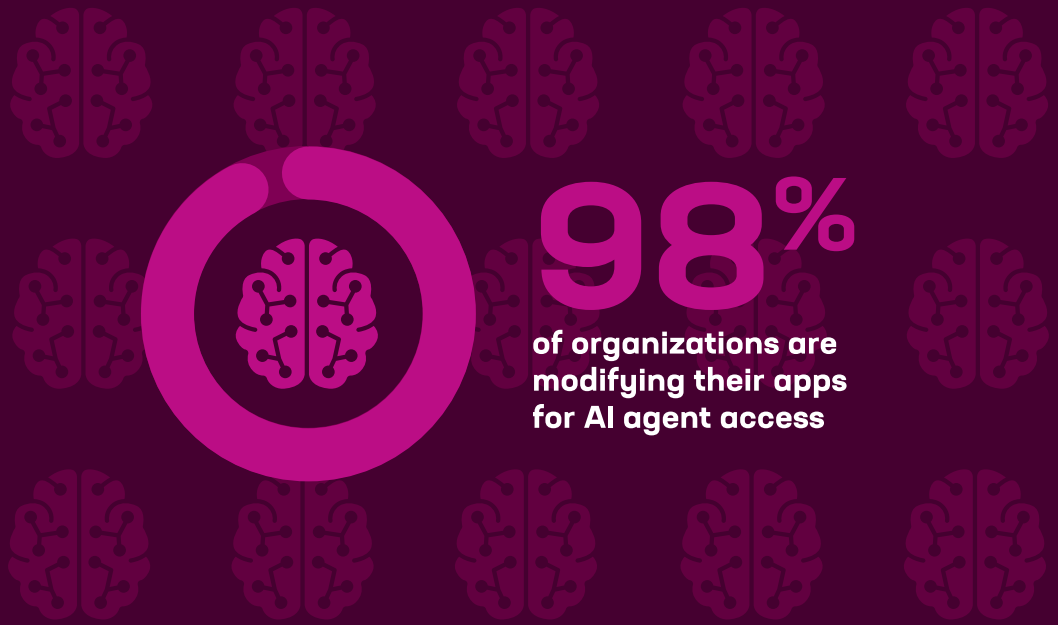
In addition to managing distributed inference services, all but 2% of organizations are actively planning to modify their external facing applications to interact with AI agents. The single most common approach is implementing identity-aware infrastructure using tools such as advanced load balancers to route and manage traffic based on machine or agent identity. Nearly half (47%) expect to implement identity awareness. One-quarter are focused on developing and documenting public-facing APIs to allow AI agents to directly access app data and functions. Another 20% are adopting semantic data standards and data enrichment to ensure AI agents can understand the context/meaning of information without ambiguity.

AI is no longer an R&D artifact. It's a production workload embedded in daily workflows and materially impacting apps and the business.

AI AGENTS DRIVE APP CHANGES

We asked:
How are you planning on modifying your external facing applications to operate with AI agents?

We learned:
98% of organizations are modifying their apps for AI agent access, and nearly half of those are implementing identity-aware infrastructure.



Observability and automation support inference

The status of AI inference as an integrated business norm is supported by our survey findings on observability and automation.

First, operational data has largely transformed from an alerting and reporting tool into a runtime dependency. Its use has jolted upward from modest levels in 2023 to percentages in the mid-80s in 2025 to near universality today.

These steep adoption curves are not characteristic of organizations dabbling in new tools. Rather, they mark wide recognition that an enterprise can't run AI—or anything modern—without feeding those systems constant, structured, contextual information. As a result, observability is no longer a mirror held up to the system. It has become part of the system itself.

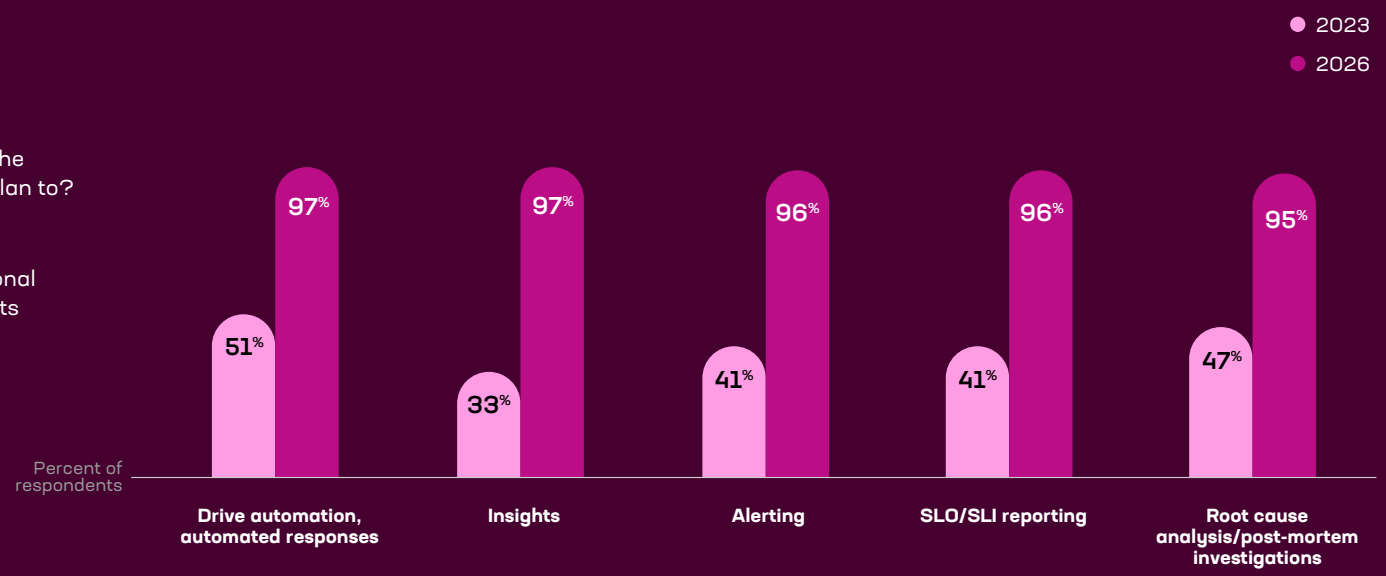
In a similar evolution, organizations today are consistently integrating AI into their automation workflows. But AI in operations is no longer confined to advisory roles. Our survey data shows a sharp transition from the use of AI primarily as a decision-support tool to its role as execution-capable actor, albeit one with clear guardrails.

Observability is no longer a mirror held up to the system. It has become part of the system itself.

OPERATIONAL DATA USE HAS SKYROCKETED

We asked:
Do you currently use operational data for the following [purposes]? If not, when do you plan to?

We learned:
Nearly everyone is currently using operational data or plans to do so soon, both for insights and to drive action.



Two-thirds of organizations already use AI to automatically adjust policies and configurations, as well as to accelerate automation efforts. That's even more than the nearly 60% of organizations relying on AI-generated recommendations to guide human action. For a majority of organizations, therefore, AI is no longer experimental. It is embedded and active in the operations control loop.

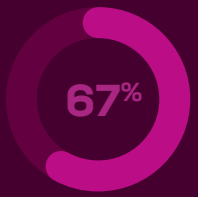
Yet our survey data indicates that fully autonomous execution is selectively limited to operational tasks whose impact is both measurable and reversible. These typically address cost control or app performance, such as traffic adjustments and optimizations to improve the user experience. People still oversee and control most app security, compliance, and business-risk decisions.

That means AI is not replacing human oversight or ownership in IT operations. Instead, it serves as a constrained actor, executing decisions to augment control loops while quietly reshaping how decision makers think about automation, observability, and the day-to-day work of running modern systems.

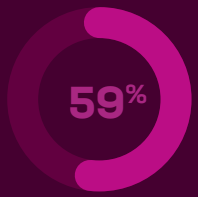
AI IS ACTING IN IT OPERATIONS

We asked:
Please select how you use AI for automation within your IT operations.

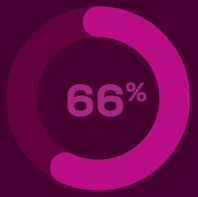
We learned:
AI is shifting from an advisor to a constrained execution engine.



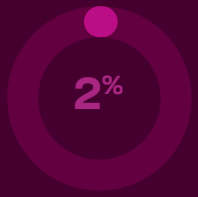
To accelerate automation efforts



To generate suggestions for adjustments to policies and configurations to meet established SLOs



To automatically adjust policies and configurations to meet established SLOs



We do not use AI for automation

Furthermore, regardless of how AI is being used within IT operations—whether for accelerating automation, generating suggestions, or adjusting policies—respondents consistently ranked programmatic interfaces as their preferred method for managing app delivery and security services. API-driven automation ranked highest across all use cases. Even respondents who reported that they do not currently use AI for automation preferred API-driven mechanisms for app service management. This focus on machine-consumable governance facilitates automated operations.

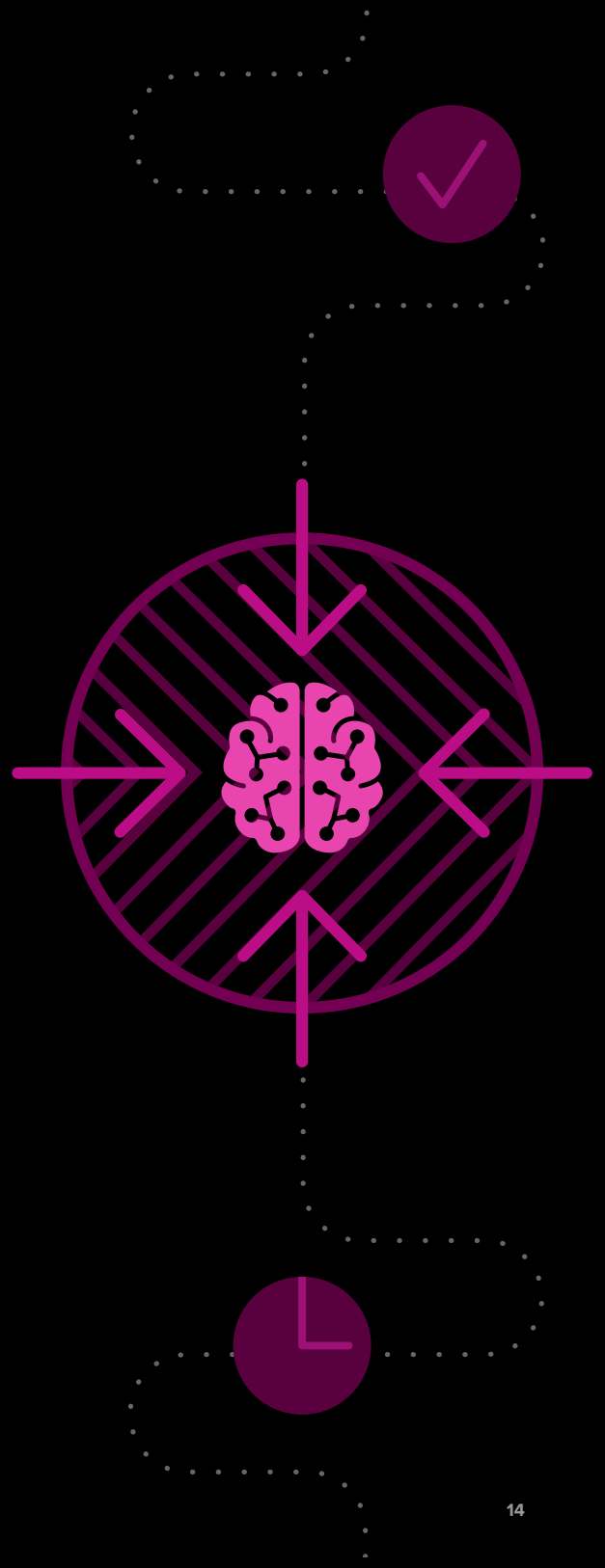
Two conclusions stand out. First, automation is no longer a sidecar to operations; it's how AI becomes practical at scale. Second, organizations are already designing for systems that act, react, and adapt without human mediation, wrapping automation around inference to make it operational. Since the depth of automation around a workload is a marker of that workload's maturity, we can conclude that inference is coming of age.

Inference weighs in as the operational—and value—center of gravity

Moreover, when asked where application delivery and security services have the greatest operational impact in AI architectures, respondents overwhelmingly pointed at the capabilities wrapped directly around inference. Respondents see the highest value around, and therefore prioritize, protecting input activities such as input filtering, injection prevention, memory merging, and prompt handling. Output moderation is also seen as important, but today's primary focus is on protecting input to avoid "garbage in garbage out" scenarios.

64%

allow AI to adjust policies or configurations



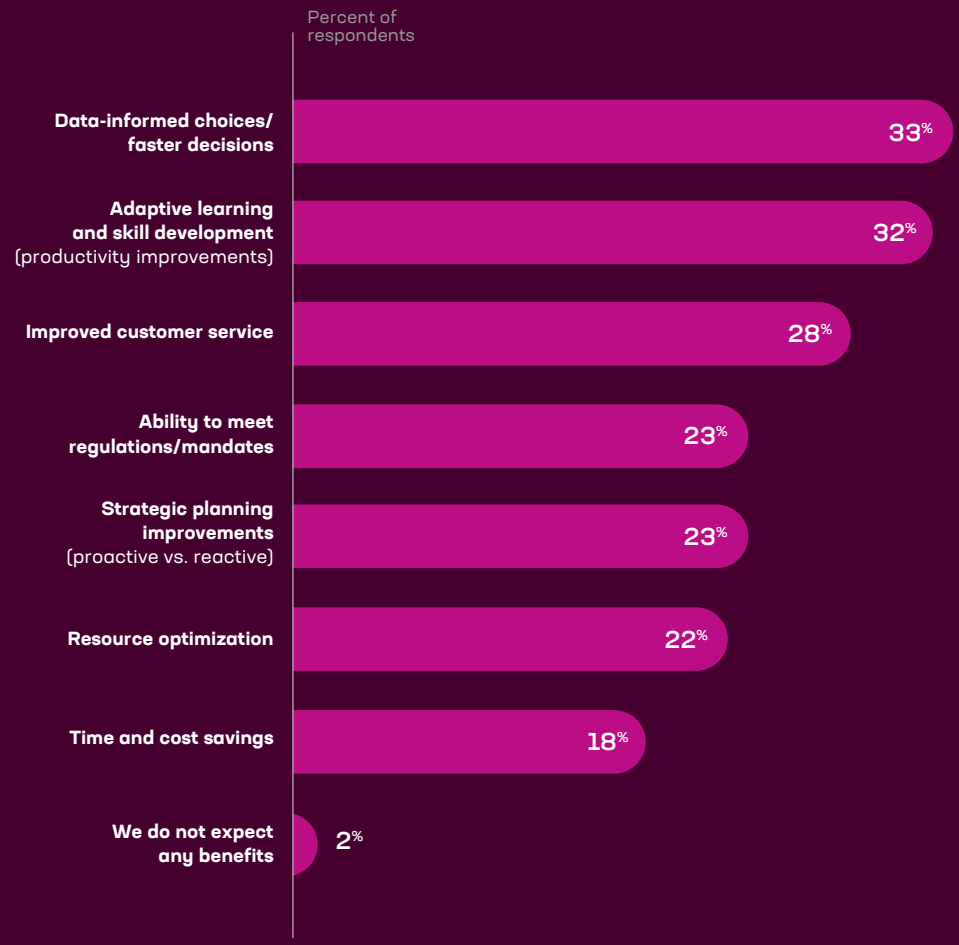
This focus is partly a function of the value organizations expect from generative AI. First, enterprises are, by and large, not chasing creativity. The top two expected outcomes from generative AI are faster decision making (33% of respondents) and productivity improvements through adaptive learning (32%). These are control-plane benefits. They depend on AI being embedded into operational workflows, decision loops, and day-to-day execution, not isolated behind a chatbot interface.

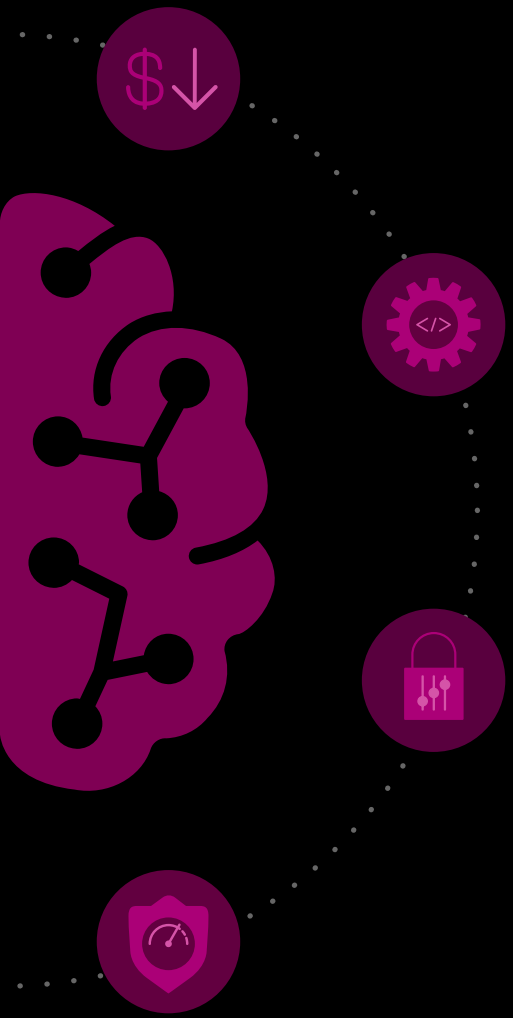
Faster decision making is the **number one value** of generative AI

SPEED AND PRODUCTIVITY TOP AI VALUES

We asked:
What are the top two benefits you expect from using generative AI?

We learned:
Faster decisions and greater efficiency narrowly beat other values.





Significantly, other potential AI benefits, including cost savings, strategic planning, and even regulatory compliance, trail the top two. Organizations expect AI to accelerate work more than they expect it to foster innovation or transform strategy. Similarly, the factors influencing model choices, such as integration compatibility, costs, and availability, are issues of resilience, not innovation.

When faster decisions and productivity gains are the primary value drivers, then controlling inputs, shaping prompts, enforcing constraints, and managing memory become the highest-leverage controls. That's why respondents report that the greatest operational impacts of AI app delivery and security services are concentrated on inputs. They're optimizing the front door of inference because they can influence those outcomes without surrendering control.

The operational heart of AI beats in inference—not in training pipelines, data labeling, or offline model design. Enterprises are treating models as interchangeable execution engines rather than the primary locus of intelligence. That locus is inference, which has become a runtime responsibility. As such it redefines application delivery and security priorities.

Value and risk are concentrated at the systems governing inference

Operational pain isn't slowing AI adoption

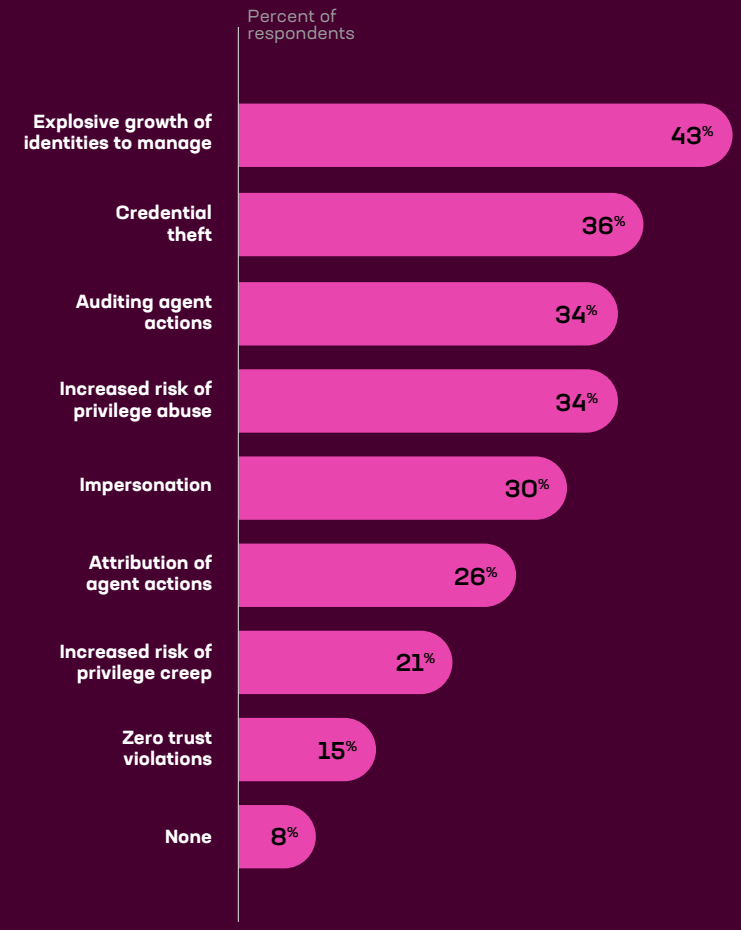
There's no question that even as AI has proliferated, its deployment requires overcoming barriers. Nearly nine in 10 respondents report at least one constraint. The cost of compute resources tops the list, and access to sufficient skills ranks a close second.

Furthermore, once those barriers are surmounted, agentic AI brings a whole new set of challenges to trouble more than nine in 10 respondents. The most common is expected to be managing explosive identity growth as AI agents proliferate. Other top concerns include credential theft and other visibility and security issues that traditional IT and security models were not designed to handle. Organizations that don't carefully manage their AI adoption risk significant friction and harm.

TEAMS WORRY ABOUT EXPLOSIVE IDENTITY GROWTH

We asked:
What challenges do you anticipate for agentic AI? Select all that apply.

We learned:
Identity management is the number 1 concern, thanks to the proliferation of AI agents. Other security worries are common.



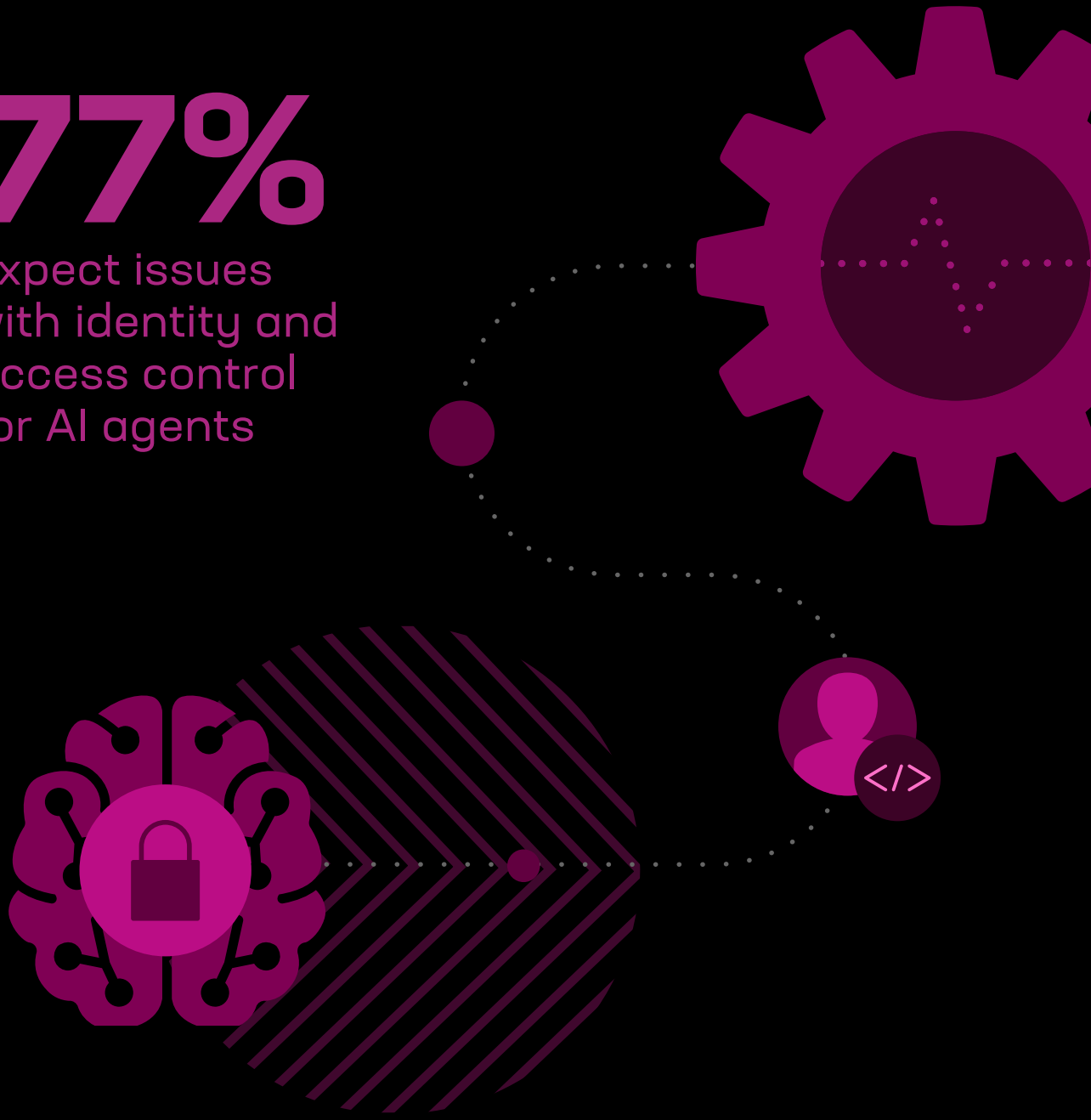
Nobody's stopping, however, and that's as true in IT operations as in strategic business functions. Across operational tasks traditionally viewed as time-consuming—such as troubleshooting, integrating with ticketing systems, and dealing with vendor APIs—the data shows no slowdown in AI adoption. Organizations have reached a level of operational confidence where inefficiencies do not suppress their AI use. Teams simply work through them because the value of AI inference outweighs the hassle.

There's also no correlation between AI adoption and app deployment environments or how they're structured. AI progress spans them. When a workload becomes essential, existing processes adapt around it. Inference has reached that milestone.

The value of AI inference outweighs the hassle

77%

expect issues with identity and access control for AI agents



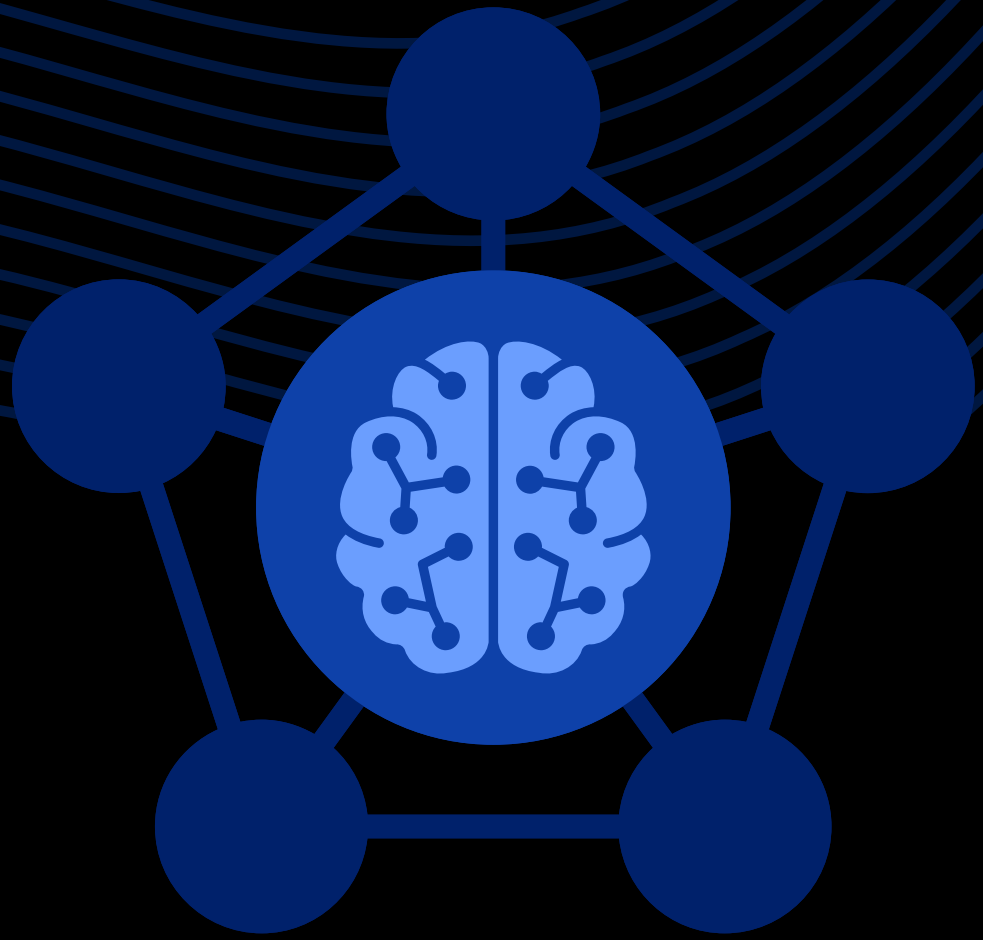
F5 INSIGHT:

Inference has joined the app stack

AI inference is no longer a flashy experiment, a hot trend, or a future concern. For many organizations, AI already sits in the daily path of business outcomes. Inference services have become an operational workload supported by mature observability, embedded automation, and well-understood architectural layers.

As a result, models are no longer the key unit of operation for AI. Instead, the system that governs how requests are shaped, routed, constrained, and observed is where value and risk are concentrated. From an architectural perspective, intelligence is moving up the stack into the delivery fabric. Organizations that have internalized this shift are treating inference services with the same gravity as any other critical component of the application stack and integrating it into their operational rhythms.

The question, then, isn't whether or how quickly organizations will operationalize AI. It's how effectively they recognize that they already have and figure out how to manage it well—without the tool sprawl, responsibility gaps, and complexity challenges that often impair other stack components. Such challenges can derail an organization by creating security risks or onerous operational costs, so evading or solving them becomes critical for an AI strategy to deliver a competitive advantage.

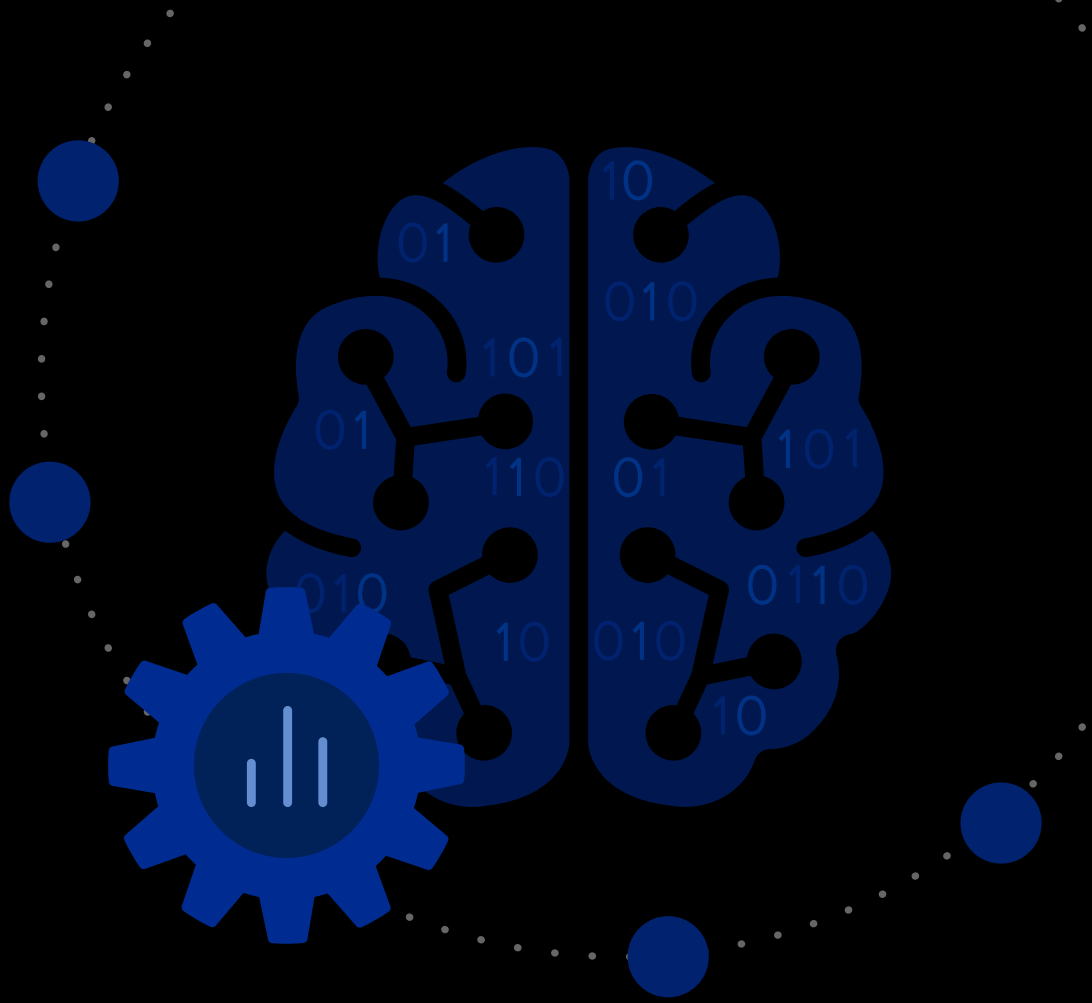


SECTION 2

Recognize AI as infrastructure

Inference looks straightforward when you only use one AI model.

You send traffic to the model, wrap it in authentication, maybe add a guardrail or two, and move on. That mental construct collapses under the evidence for what's happening today in most organizations. Our survey data shows that most organizations rely on more than one AI model. The use of five or more different AI models is common, and organizations are operating or actively evaluating an average of seven. That number invalidates the idea of inference as a single endpoint. AI is no longer about choosing the best model. It is about managing a portfolio of models and services.



Business considerations drive multi-model complexity

This multi-model approach has become the norm not because no single model offers sufficient sophistication or because the teams involved have nothing better to do than experiment with each shiny toy. Instead, multi-model AI is driven by business and technical concerns. Different models may invoke different costs, expose incompatible interfaces, and fail in different ways under load. A single-model strategy cannot optimally satisfy every need.

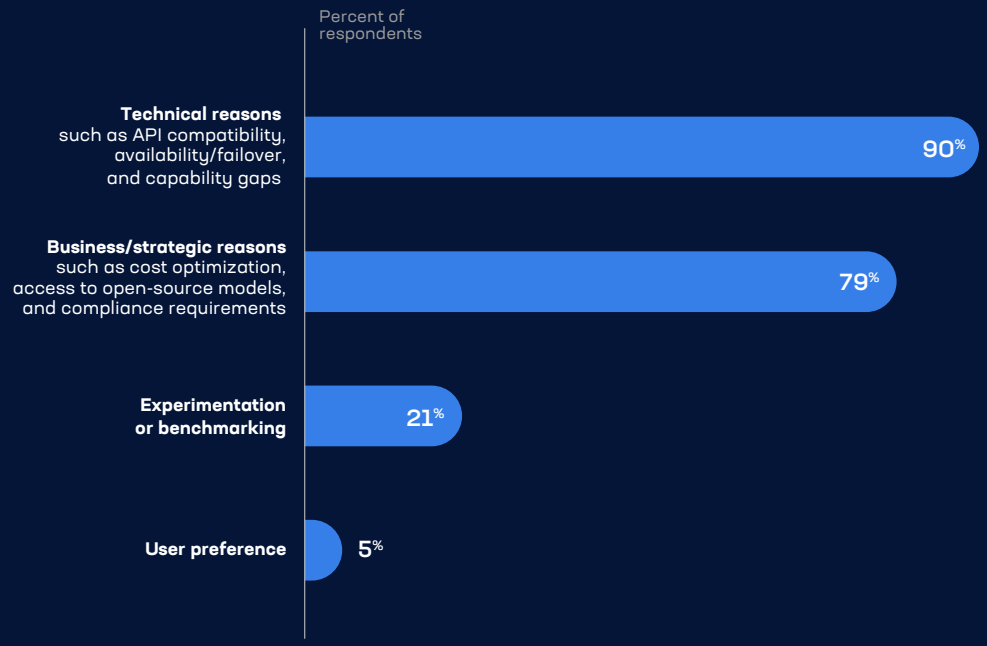
When deciding on multiple AI models, organizations consider many variables related to business and strategic needs. These include cost optimization, compliance requirements, and strategic diversification.

Technical considerations such as resiliency and specific model capabilities also come into play. For instance, organizations must ensure API compatibility to avoid app disruption caused by model lock-in or the need to rework apps to interface with a particular model. Like hybrid multicloud app deployment, multi-model AI architectures respond to the reality that no one model excels at every task or suffices for every data class or jurisdiction.

MANY CONSIDERATIONS DRIVE MULTI-MODEL AI

We asked:
What are the top three reasons your organization uses different AI model families?

We learned:
Multi-model AI exists for the same reasons as hybrid multicloud app deployments–because business and technical strategies demand it.



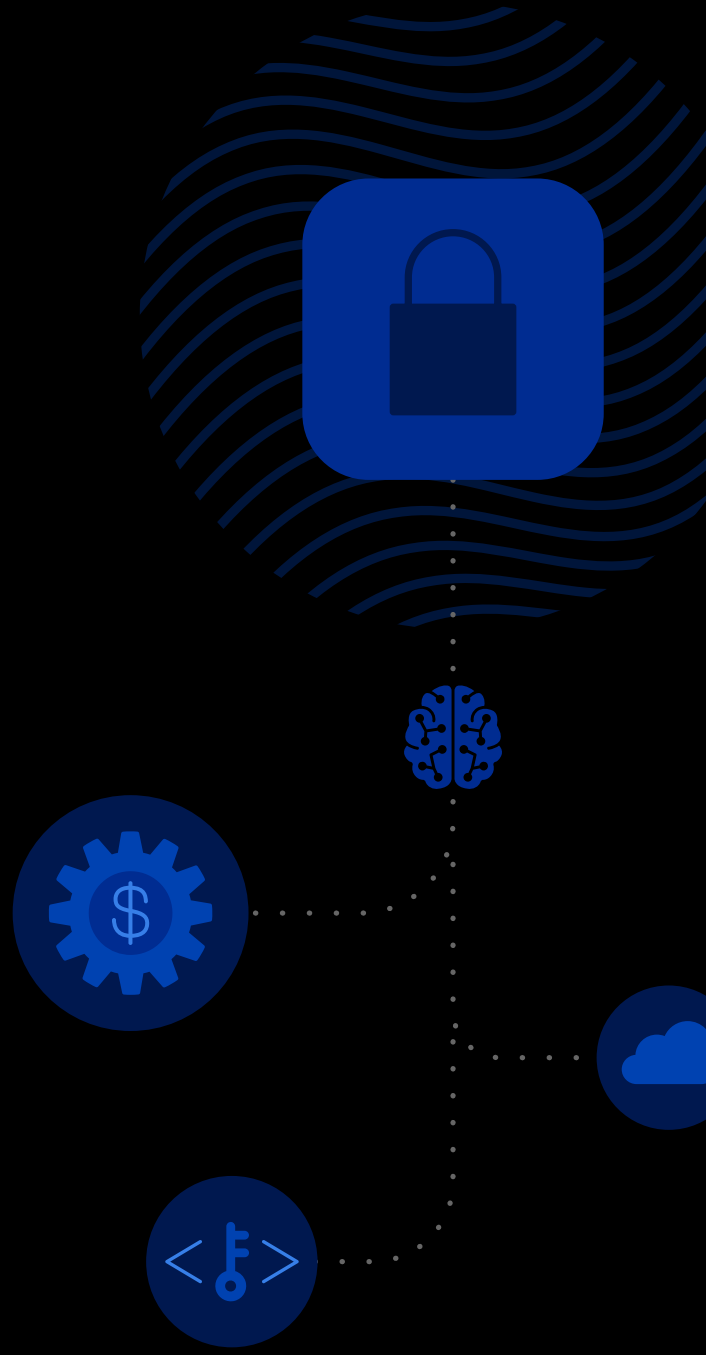
When inference becomes part of a production experience, competitive enterprises need to deliver AI while preserving existing app integrations, managing per-request costs, and maintaining availability under failure conditions. Multi-model inferencing allows enterprises to route requests based on function, sensitivity, and policy, treating inference traffic the same way they already treat application traffic in regulated environments.

The ranking of strategic diversification and access to open-source models as justifications for multiple models further demonstrates that multi-model AI amounts to risk management, not experimentation. Notably, user preference barely registers as a driver. Organizations are not choosing multiple models based on data science preferences but because the enterprise architecture and business imperatives demand it.

Once multiple models exist in parallel, inference stops behaving like a stateless service and starts behaving like a distributed system in which different models serve different architectural roles. Some may be large and general purpose, others narrow and tuned for a specific domain. Some may be optimized for cost and throughput, others for accuracy and business value. The question becomes one of routing: which model should handle a given request, under current conditions, without violating latency, cost, security, or compliance constraints. The criteria for such routing decisions are factors such as API compatibility, cost optimization, security, and availability.

52%

of organizations are chaining or orchestrating multiple models



Our survey data about how AI models are being adapted demonstrates this shift to a distributed inference system. More than half of organizations are manipulating models. Most of the adaptation techniques in use are lightweight or represent small-scale tuning. But multi-model chaining and orchestration leads all others, outpacing even prompt engineering. More than half (52%) of organizations use multi-model chaining or orchestration. Nearly as many use knowledge distillation, and more than one-third use prompt engineering or prompt templates. Retrieval-augmented generation (RAG), often treated as the default pattern in architectural diagrams, trails significantly.

This result shows that practitioners are investing first in mechanisms that enable them to coordinate, compress, and route across models rather than simply enriching a single model with more data. In other words, they are designing systems, not tuning artifacts.

Infrastructure implications include the need for orchestration

The systemization of AI has direct implications for the enterprise architecture. When organizations distill large models into smaller "students," pair them together, or dynamically chain models based on task or tenant, the focus shifts from the models themselves to the control plane that decides where inference traffic goes, why, and how to protect it.

The operational behaviors that follow look very familiar from an infrastructure perspective: routing, fallback logic, versioning, cost-aware selection, protection, and risk-adjusted decision flows. Orchestrating multiple models turns inference into a managed workload subject to delivery, security, cost control, and resilience concerns. The big decisions become how to design and manage the systems that govern how inference traffic is shaped, routed, constrained, secured, and observed. Familiar operational pressures reappear. Cost becomes variable at request time rather than fixed at deployment time. Latency fluctuates based on GPU availability, queue depth, and model size. Failures become conditional, triggered by drift, degraded performance, or breached policy thresholds.

These are classic application delivery and security concerns resurfacing in a new domain. In fact, enterprises are treating distributed inference as a new application tier. They must, since nearly everyone reports that they are self-hosting inference infrastructure rather than fully outsourcing it.

Almost nine in 10 survey respondents (88%) are already using at least one strategy to deliver and secure AI inference services or plan to do so within the year.

More than half report managing authentication and API access, monitoring AI traffic flow, or preventing outbound data leakage. More than one-third report a centralized point of control such as load balancing for AI traffic across models and back-end systems. These techniques map cleanly to long-standing app delivery and security responsibilities. In addition, monitoring public cloud costs for AI workloads is now intertwined with AI security and availability because those inference decisions directly impact expenditures and business risk.

88% have deployed at least one app service for inferencing

In short, enterprises are operating AI systems with all the same performance, availability, and security demands as their application delivery and security systems.

INFERENCE REQUIRES TYPICAL DELIVERY AND SECURITY SERVICES

We asked:
How do you deliver and secure, or plan to deliver and secure, AI inferencing services? Select all that apply.

We learned:
The vast majority (88%) of organizations have deployed at least one service for inferencing delivery and security, and more than half use multiple services.

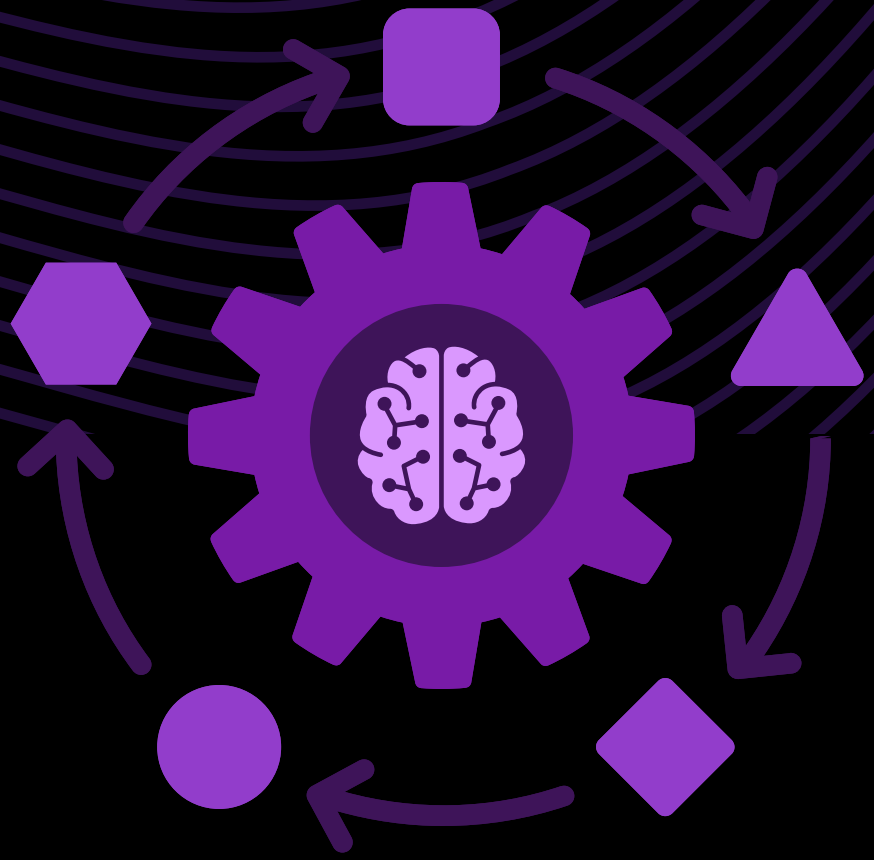
F5 INSIGHT:

AI systems need an orchestrated control plane, and app services can play that role

In a multi-model AI world, the routing logic, policy enforcement, protection, observability, and feedback loops around inference matter more than individual model choice. Yet many organizations still describe their AI strategy in terms of models rather than systems. This creates a dangerous mismatch between reality and governance.

We've seen the risks of such a mismatch before. Enterprise architectures underwent a similar transition when monolithic applications gave way to microservices and APIs. What followed was the rise of load balancers, gateways, service meshes, and centralized policy engines. AI inference is on the same path to a distributed system—and to the same risk of burdensome complexity and security risks if that system is not managed well.

The difference between the two transitions is speed. When it comes to AI, organizations have already moved into fleet management mode. Some will discover the hard way that running seven models without orchestration is indistinguishable from running seven production systems with no traffic management and inconsistent security. Digital leaders will recognize early that inference is essentially an application delivery and security problem, and from that perspective, they will build control planes that can scale as AI complexity grows. The more they treat these systems like a layer of architecture, the more effectively they're likely to manage them.



CONCLUSION

Proactively manage and secure AI infrastructure via app services

Taken together, our 2026 survey results reveal a digital community that is operationally pragmatic and aware of the risks of unfettered growth in AI deployment.

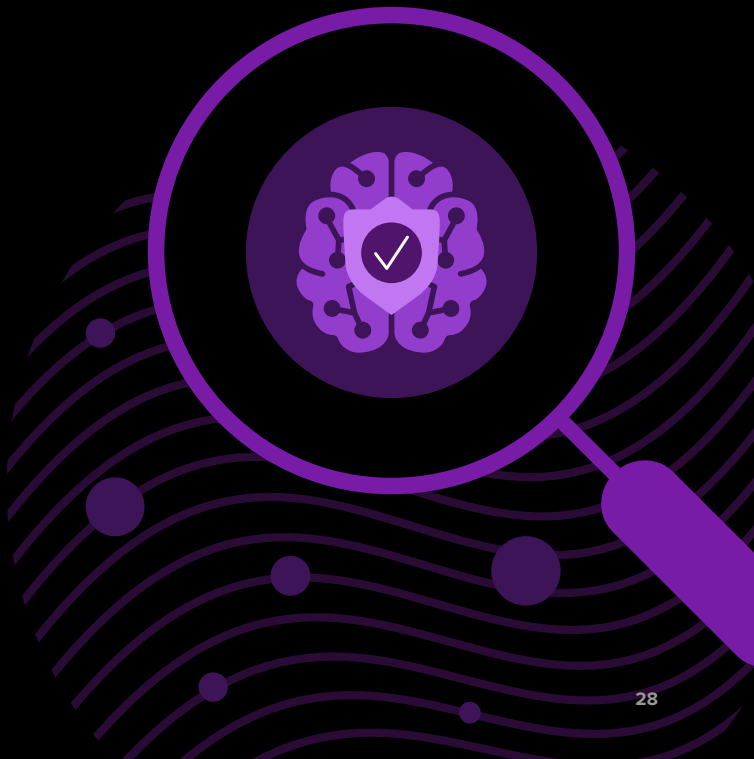
They want AI to help people decide faster and work better, not to replace their judgment entirely. To that end, IT organizations are comfortable letting AI tune operational systems continuously, but only within clearly defined boundaries.

To achieve those goals, organizations are coordinating multiple AI models and multiple inference services to optimize for business considerations ranging from availability to zones of jurisdiction. Based on those same considerations, they're also investing in delivery and security controls and applying them, especially to intent and context in the prompt layer, where guardrails can be erected without slowing the business. These investments are turning inference into a managed, policy-driven workload.

For application delivery and security teams, the critical work is no longer choosing or training an AI model but designing the systems, policies, and limitations that determine where AI can act, when it must defer, and how its actions and output will be observed, constrained, protected, and potentially reversed or corrected. That work is not an AI problem. It's an enterprise architecture problem.

To achieve the AI values they prioritize without getting completely bogged down in managing the systems that deliver those values, decision makers need to think ahead about how to consolidate their AI system control. Organizations that recognize the nature of this problem early will operate architectures that treat AI like any other critical production system, with strong, converged control planes and shared delivery and security services.

In that way, they will ensure success for their AI projects. Others will stall under their own caution or succumb to complexities that slow deployment and production systems, increase costs, invoke risks, and stymie scalability. Only the digital leaders will build architectures able to safely scale AI and capture its benefits at scale, too.

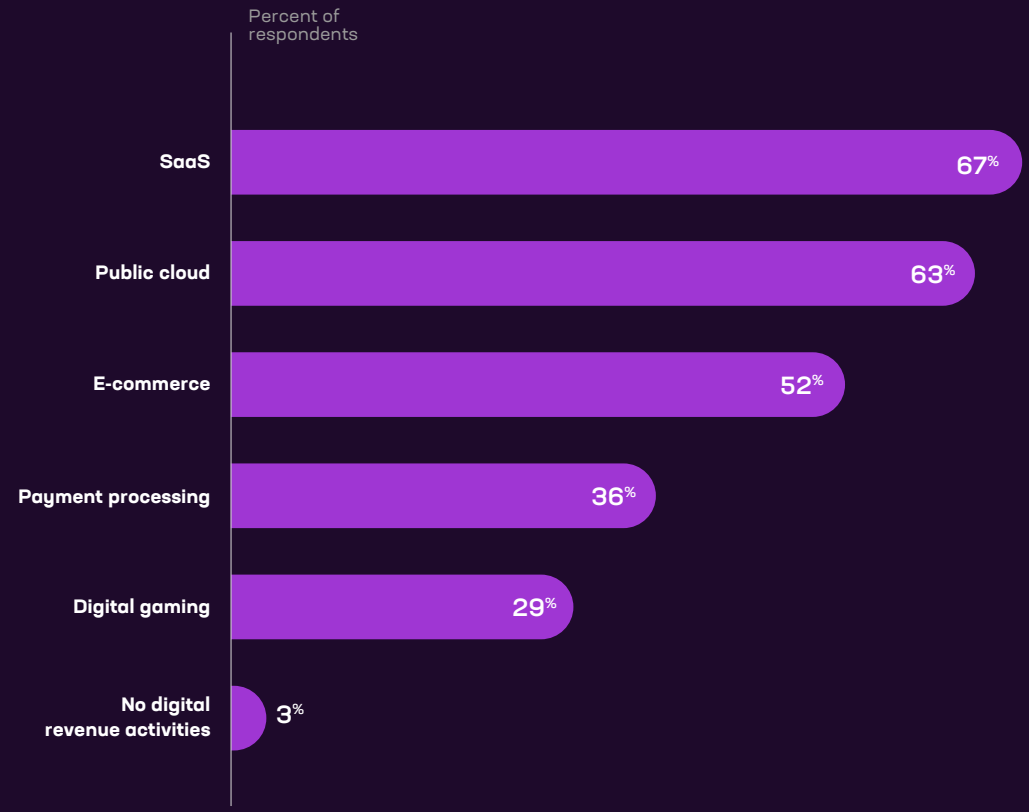


About this report

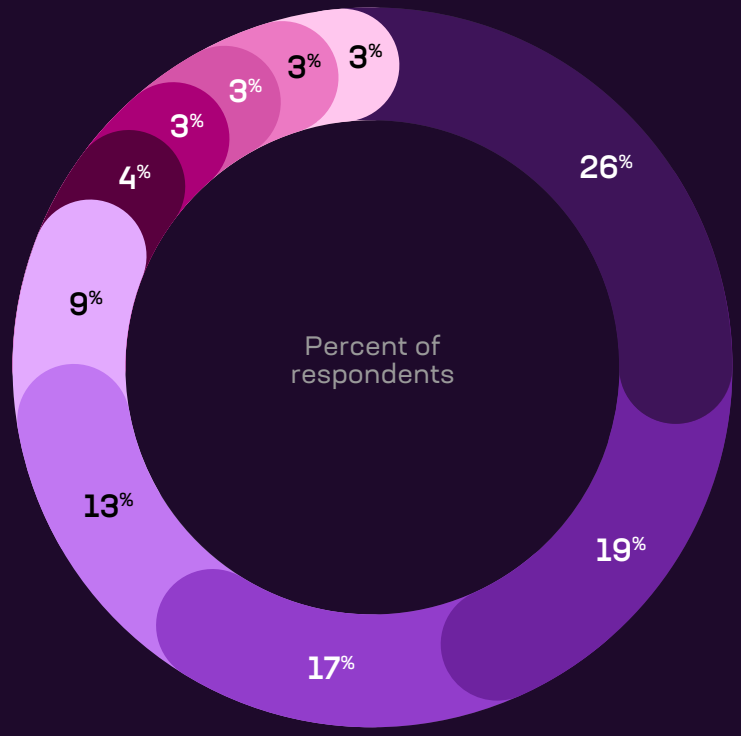
More than 1,100 IT decision makers from around the globe completed the 2026 F5 State of Application Strategy survey, our twelfth annual. Three-quarters of respondents represented organizations earning between $1 billion and $10 billion USD (or the equivalent) in annual revenues. This weighting, a departure from previous years, was intentional. After more than a decade of F5 surveys, this market segment has most consistently anticipated future trends in application delivery and security.

The revenues reported by respondents were generated at least in part by digital activities ranging from ecommerce to SaaS. Only 3% of respondents work for organizations (such as some government agencies) that do not generate revenue through digital applications.

DIGITAL REVENUE GENERATION



Respondents represented a variety of market sectors, with the cloud, technology, and financial services industries particularly well represented. Two-thirds (67%) were IT decision makers with budget authority, and more than one-third (39%) were IT executives at the C-level, such as chief information or chief technology officers.



ABOUT F5

F5, Inc. (NASDAQ: FFIV) is the global leader that delivers and secures every app. Backed by three decades of expertise, F5 has built the industry's premier platform—F5 Application Delivery and Security Platform (ADSP)—to deliver and secure every app, every API, anywhere: on-premises, in the cloud, at the edge, and across hybrid multicloud environments. F5 is committed to innovating and partnering with the world's largest and most advanced organizations to deliver fast, available, and secure digital experiences. Together, we help each other thrive and bring a better digital world to life.

For more information, go to f5.com.