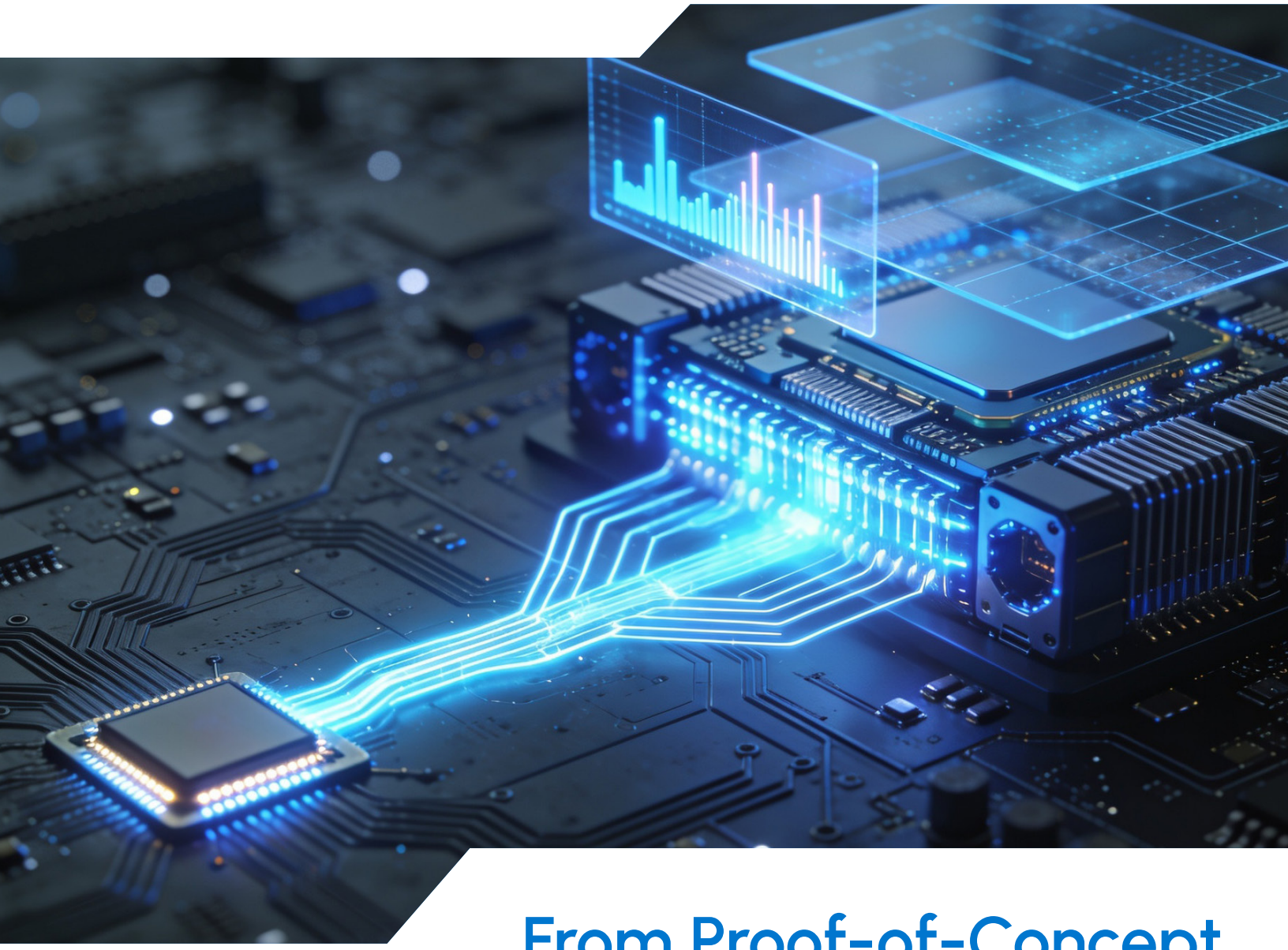
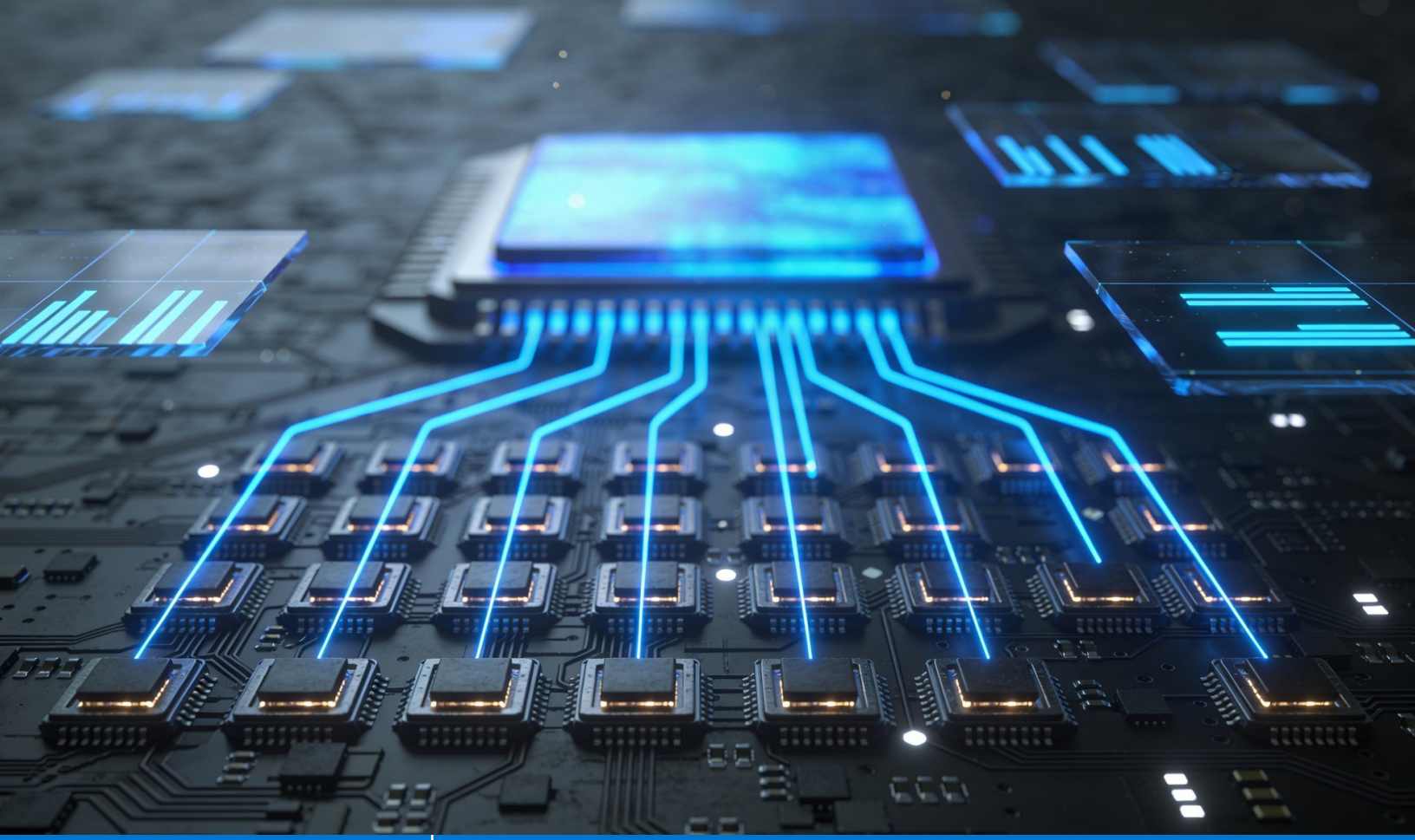
Futurum RESEARCH

DELL Technologies



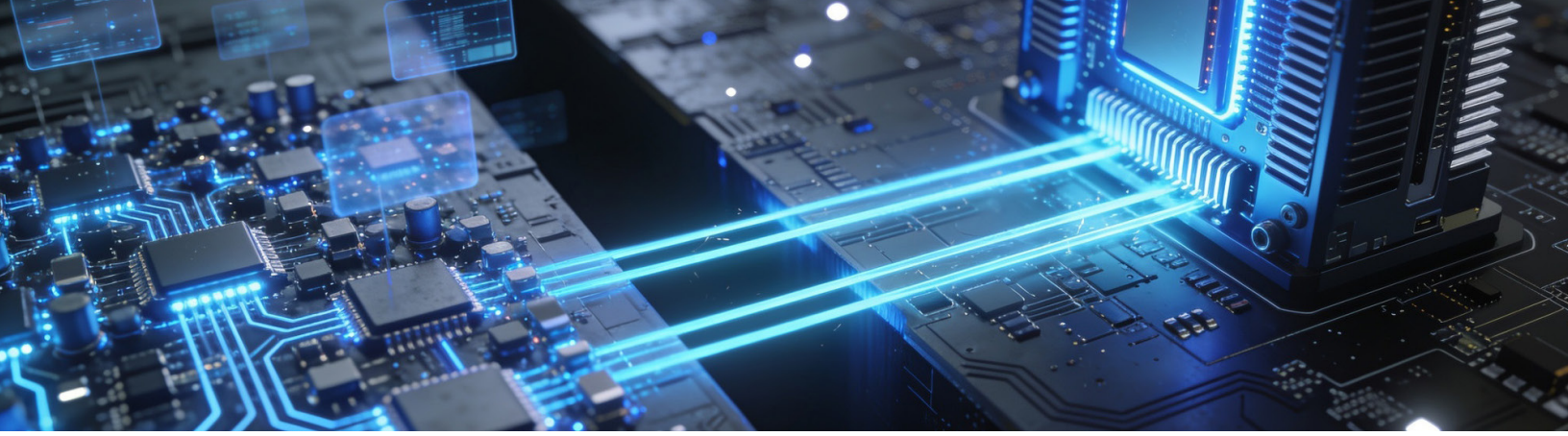
From Proof-of-Concept to Value: Building an AI Engine for Production

NOVEMBER 2026



In brief:

Most enterprises can now prove that AI works; far fewer can run it at scale. Futurum survey data shows more use cases in pilot than in production, with momentum lost at the handovers between platform, workload, data, governance, and security teams. This paper makes the case for an integrated, production-first operating model, including the approach Dell Services takes, in which cost, governance, and reuse are designed in from the first decision.

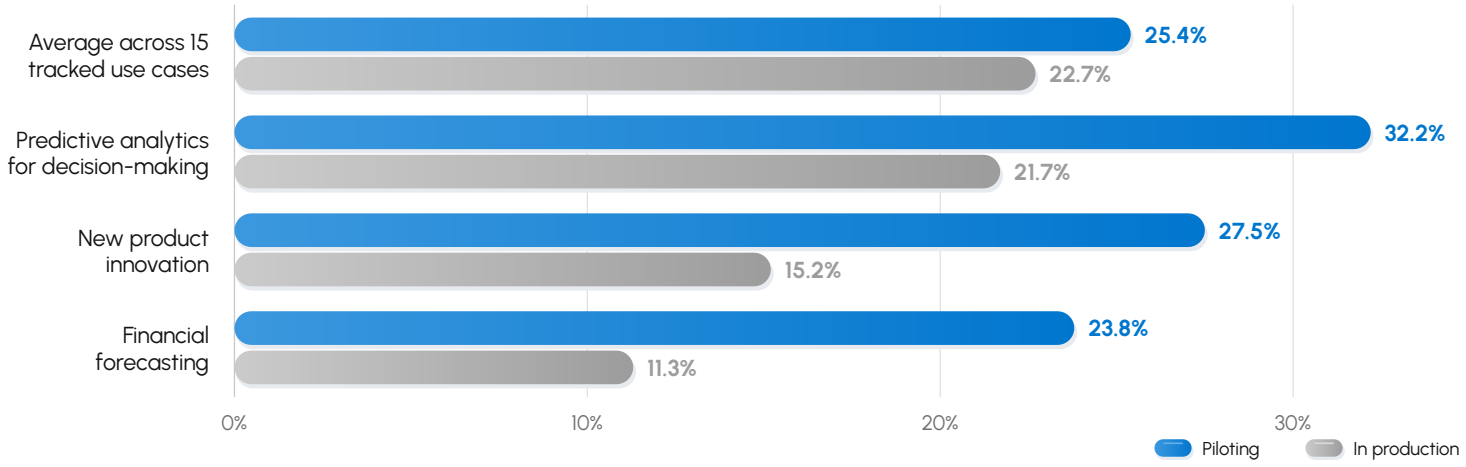


The production gap is real, and it is measurable

Enterprise AI now has the opposite of an ambition problem. Boards have settled the question of whether AI matters, budgets have been approved, and most large organizations can point to a portfolio of experiments that demonstrably work. The difficulty arrives afterward, on the road from something that works once to something on which the business can run.

Futurum Research AI Product Series data puts numbers against that difficulty. In a March 2026 survey of 600 technology leaders building AI applications internally, an average of 25.4% of the 15 tracked use cases were being piloted against 22.7% running in production.[1] More of the work sits in pilot than has crossed over, and the gap is widest in exactly the places where expectations run highest: predictive analytics for decision-making, piloted by 32.2% and in production for 21.7%; new product innovation, 27.5% against 15.2%; financial forecasting, 23.8% against 11.3% (see Figure 1).

Figure 1. Use Cases Where Pilots Most Outpace Production



Note. Percentage of respondents piloting versus running each use case in production; the average is across the 15 use cases tracked. N = 600. Data from The Futurum Group (2026), AI Product Series, Theme 3: Building AI Internally, March 2026 survey.

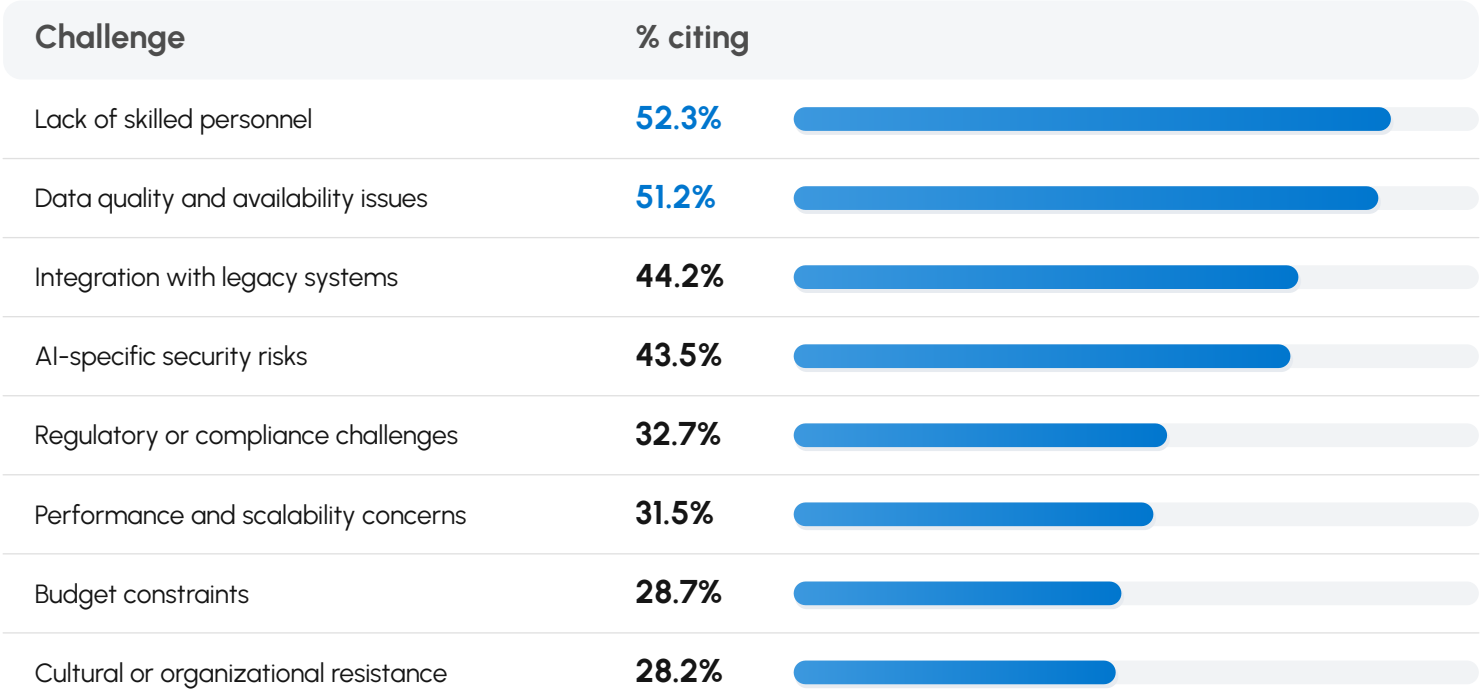
Those ratios show an industry that has figured out how to prove its value but has not yet figured out how to scale it. The use cases that have broken through are the contained ones, with employee support chatbots at 39.2% in production and enterprise search at 34.5%, both of which have climbed in each of the three survey waves since March 2025. The work that changes how a business makes decisions or prices its products trails well behind. Pilot fatigue is the accumulated weight of that lag, and it is a governance and operating-model problem long before it is a technology one.

1 The Futurum Group. (2026, March). AI Product Series, theme 3: Building AI internally [Unpublished survey data; N = 600; n = 467 for production-only questions].

Where momentum actually breaks down

The instinct when a program stalls is to look for the component that failed, often the infrastructure or the model. The evidence points somewhere less convenient. Asked what gets in the way of deploying use cases in production, the 600 technology leaders in the March 2026 wave spread their answers across the entire stack, with no single factor dominating (see Figure 2).

Figure 2. Challenges to Building Internal AI Applications

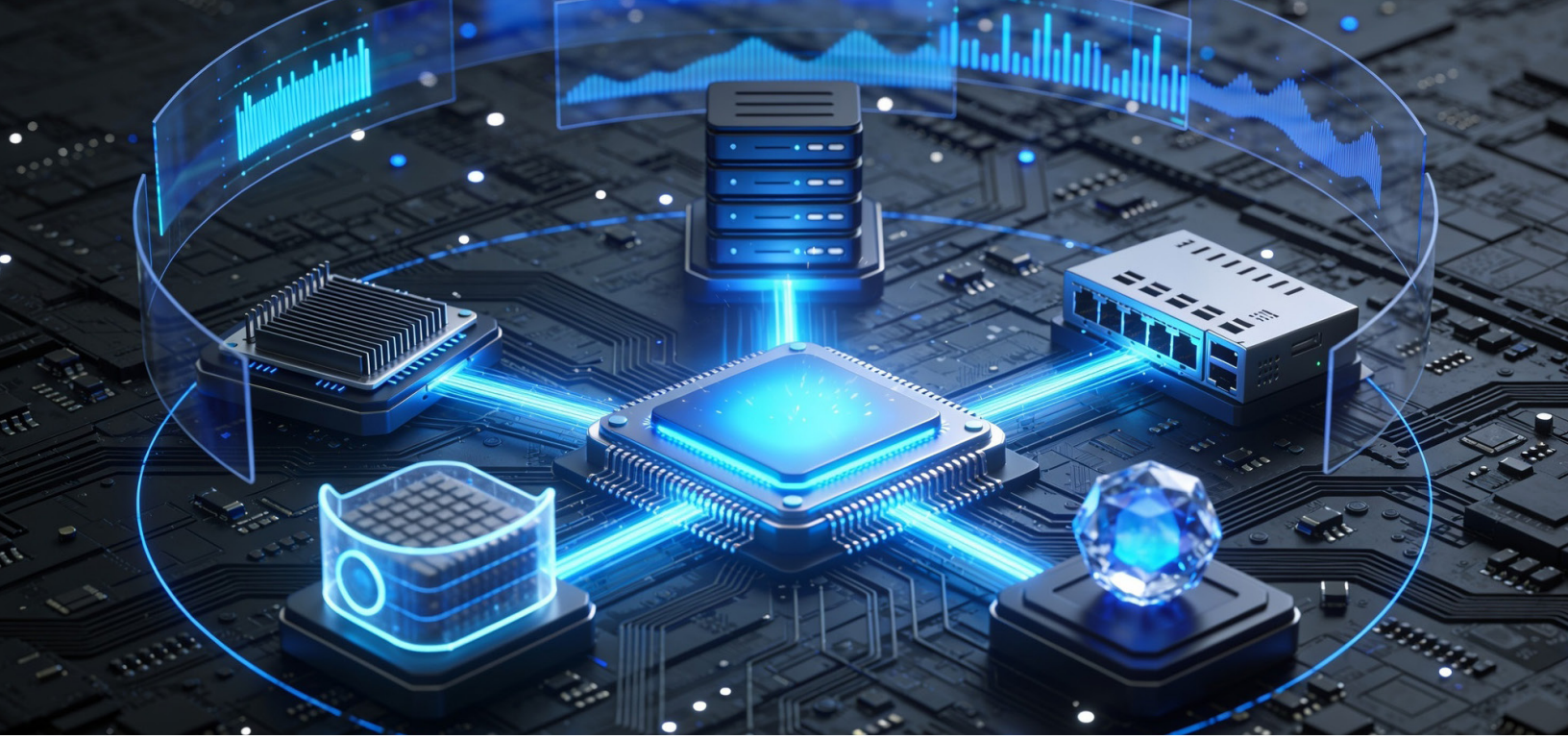


Note. Percentage of respondents citing each challenge; multiple responses allowed. N = 600. Data from The Futurum Group (2026), AI Product Series, Theme 3: Building AI Internally, March 2026 survey.

The shape of that distribution matters more than any one row in it. Dell Technologies frames this as five disciplines that must work as a single production-grade system: the platform, the AI workload, the data, governance, and security. Each has its own owners, tooling, and requirements, and almost no organization holds all five in-house at the same level of maturity. Skills and data sit at the top of the list, but a program that solves both still has integration, security, and compliance behind it, including how new systems and existing infrastructure will work together.

The result is a sequencing problem. Work traditionally moves down a relay: one team completes its portion, hands the project over, and the next team re-scopes what it inherited. Every handover is a moment when momentum is lost, and losses compound. Leaders consistently underestimate how difficult those handovers are, and the misreading is expensive. It sends organizations hiring or buying to fill what looks like a missing skill, when the real constraint is the distance between two capabilities they already have.

Agentic workloads make this harder in ways that the survey data is only beginning to capture. A production agent is not a single model to govern but a set of components that take actions, call tools, and invoke sub-agents, thereby broadening the governance surface and multiplying the points at which the five disciplines must agree.



An operating model built for production

If the blockage is structural, the remedy has to be structural too, which requires a shift in focus from what decisions are made to when they are made. This is also where AI programs part company with conventional IT projects. A traditional deployment can have its requirements specified up front and verified at the end, whereas the value of AI depends on how data and model behavior hold up once something is running, which only becomes visible after the build is underway.

Conventional delivery front-loads the commitments that are hardest to reverse. Organizations are asked to pick a platform, size an environment, and lock an architecture at the point when they know least about what the business actually needs, and those early choices turn out to be remarkably sticky. A design that made sense on day one is still being built on day ninety, whether or not it still makes sense.

The better approach reduces the number of premature decisions and brings in multidisciplinary expertise before the architecture is fixed, so the choices that do get made are made at the right moment and by people who have made them before. Governance and security belong in that early group. Bolted on after a build, they force a rework of the whole solution; when shaped from the start, they become part of the foundation, enabling organizations to scale with confidence.

Dell Services answers this with a codified engineering delivery model aligned with the Dell AI Factory, which puts all five disciplines into a single accountable team that works in parallel rather than in sequence. The mechanism is straightforward, and that is rather the point, because it targets the handovers directly rather than adding more specialists at either end. There are some recent Dell Services engagements that illustrate this pattern. One example focuses on Physical AI, with a large global technology company that sought to move beyond isolated experiments by quickly advancing digital twins while maintaining tight control over data and security. What made the difference was an integrated approach, spanning compute, storage, networking, software, and services, resulting in faster time to value, including compressing building design from months to hours. Another example centers on the use of AI in higher education. A large US research university aimed to strengthen research capabilities and research status while simultaneously enriching student and faculty experiences through responsible AI use campus-wide. Instead of starting with technology choices, the approach centered on prioritizing use cases and the operating model to support them. The result was a secure research computing platform and a governance model that its own teams run.

Forward-deployed engineering is emerging as a significant shift in how vendors, integrators, and consultancies are addressing the complexities of AI, an approach that centers on placing specialized expertise closer to the customer's environment. However, not all approaches are the same. While some primarily provide staff augmentation, others focus on straightforward software deployment, leaving organizations to figure out the rest. What separates them is whether the method has been run before, and whether the engagement leaves the organization better equipped to sustain and scale what has been built.

Dell Services positions a delivery model codified from years of engagements and anchored on the Dell AI Factory, drawing on lessons from its internal agentic AI work as much as on more than 6,500 AI Factory deployments since 2024.[2] Its engineers work the five disciplines in parallel rather than in sequence, fitting the build to each organization's existing constraints and security and governance requirements, with skills transfer running throughout the engagement so the customer's team can operate and evolve the environment afterward. Dell puts the resulting time to value at 84% faster than do-it-yourself deployment.[3] For organizations that would rather not run it themselves, there are fully managed operations, including proactive monitoring, lifecycle management, and showback reporting for AI spend, backed by a 99.99% availability SLA.[4]

Ask any vendor making this pitch, Dell included, who is operating the environment twelve months after handover, and how many of the customer's own engineers can change it without the vendor in the room. The answers vary far more than the marketing does.



2 Dell Technologies. (2026, August). Analysis of customer order data [Unpublished internal analysis]. Actual results may vary.

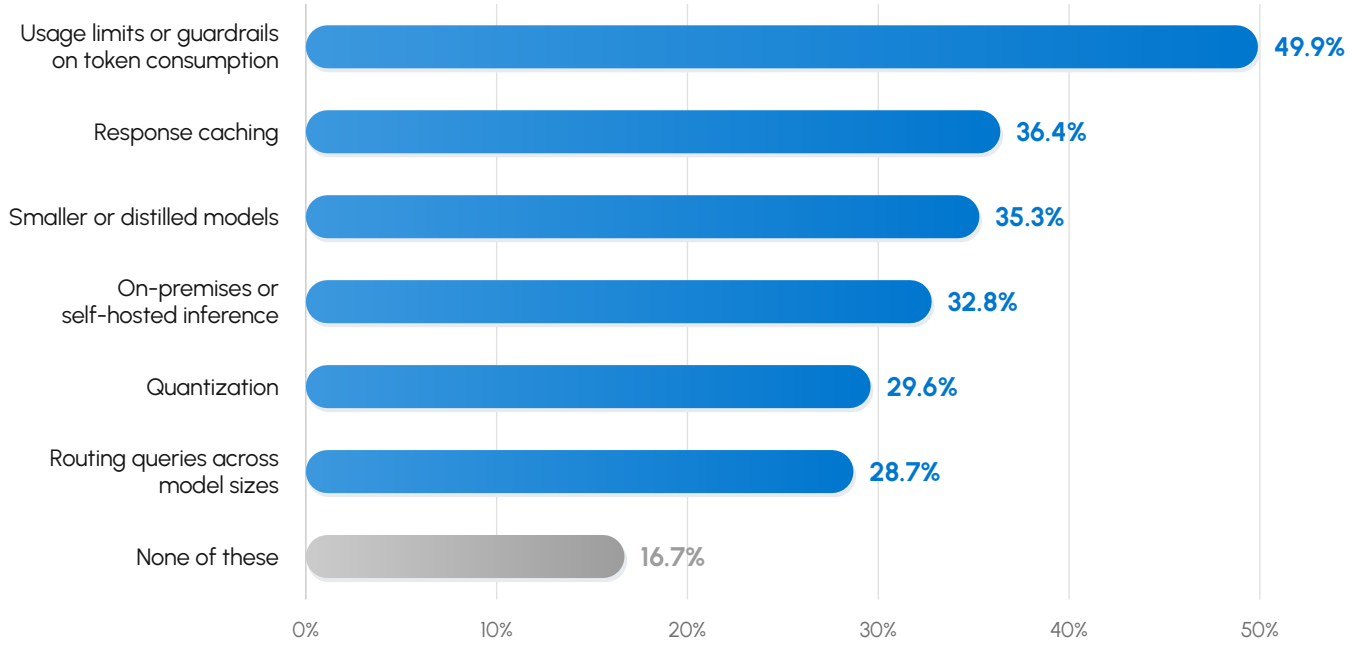
3 Principled Technologies. (2026, May). Accelerate AI time to value with Dell Services [Report commissioned by Dell Technologies]. https://www.delltechnologies.com/asset/en-us/services/deployment/industry-market/principled-technologies-accelerate-time-to-value-ai-factory-deployment.pdf.external Testing reflected an air-cooled single-rack 16-GPU AI Factory solution. Actual results may vary.

4 Dell Technologies. (n.d.). Dell Managed Services statement of work (SOW). Dell Managed Services include a 99.99% uptime Service Level Agreement (SLA) to ensure high availability for critical infrastructure. This SLA covers infrastructure availability, with service credits provided if downtime exceeds the specified thresholds. For detailed terms, conditions, and exclusions, refer to the Dell Managed Services SOW.

The economics of running AI, not launching it

Cost is no longer only an annual budget conversation. Futurum Research data shows how far it has pushed into day-to-day engineering decisions: among the 467 respondents currently running AI applications in production, 83.3% are actively managing inference costs in some way. Half have imposed usage limits or guardrails on token consumption (49.9%); 36.4% cache responses; 35.3% have moved to smaller or distilled models; 32.8% run inference on-premises or self-hosted; 29.6% use quantization; and 28.7% route queries across model sizes. Only 16.7% are doing none of this (see Figure 3).

Figure 3. How Production AI Adopters Manage Inference Costs



Note. Respondents currently running AI applications in production; multiple responses allowed. n = 467. Data from The Futurum Group (2026), AI Product Series, Theme 3: Building AI Internally, March 2026 survey.

These organizations are already feeling the cost pressure of running AI in production, so they are taking steps to control their spending today. But they may not yet be designing their AI systems for long-term cost efficiency from the outset. Tokens are the unit of work in an AI system, and every read, reasoning step, and response spends them. Agentic workflows multiply token consumption per outcome, because chained prompts, tool calls, and sub-agents each carry their own token cost. Small inefficiencies at pilot scale become material at production scale, which is why the useful measure of an AI program is its running cost over time rather than its launch cost.

Owning the infrastructure is one way to get control over costs, starting with knowing where the money actually goes. It means being able to see what drives spend as operating costs rise, and keeping architectural control over how the system evolves, rather than watching a metered bill climb with limited insight into why. A Signal65 study commissioned by Dell and NVIDIA found up to an 87% reduction in spend over two years for agentic AI workloads running on the Dell AI Factory with NVIDIA, compared with cloud-based APIs.[5] That figure is modeled on specific configurations and workloads, and does not substitute for an organization's own comparison. With newly announced FinOps capabilities from Dell Managed Services, enterprises gain greater visibility into GPU and token consumption, enabling them to maintain financial control and manage the economics of AI deployments as they scale.

The more durable economic argument has nothing to do with any particular deployment model. A first production use case that leaves behind reusable data pipelines, governance patterns, and platform choices makes the second one cheaper and the third even cheaper. Value compounds, or it resets. Most organizations are still paying the full cost of the first use case every time.

Economics is only part of the case; control matters as much. Where data lives, how it is governed, and where AI runs all shape how far a program can scale, and for many organizations, on-premises infrastructure makes those answers more stable and predictable. Treated this way, sovereign AI enables innovation rather than constraining it, especially when paired with building AI-native skills in-house to sustain momentum beyond the first use case.

What this means for technology leaders

The question most organizations are facing is whether they can build a path to AI value that they can walk more than once. Technology leaders need to consider not just investment in AI technology, but even more importantly, investment in skill-building. Particularly as agentic AI reshapes service delivery, organizations must build deeper internal capabilities to govern, adapt, and scale with confidence. Five things follow from the evidence above.

- **Design for production from the first decision**. Speed to pilot is the wrong metric because a fast demo that takes another year to reach production ddoes not save anyone anything. Optimize for speed to production, where value and cost both become real.

- **Put governance and security in the design, not the review**. Deploying something is not the same as having something to run the business on. That requires maintaining data control and embedding security, resilience, and governance from the start—AI-specific security risks and regulatory challenges both feature among the blockers cited above—because they cost far more to retrofit than to build in.

- **Audit the handovers and interconnectivity, not just the skills**. Map where work passes between platform, workload, data, governance, and security teams, and determine how these disparate pieces will work together as one cohesive solution. That is what keeps time from being lost at each boundary.

- **Treat cost as an architectural property**. With 83.3% of production adopters already managing inference spend, token efficiency is a design decision made at build time, not a procurement exercise conducted afterward.

- **Judge the first use case by what it leaves behind**. Pick one whose pipelines, governance patterns, and platform choices the next three can reuse, and treat reusability as a selection criterion rather than a happy accident.

Two things are worth watching over the next several quarters. The first is whether the forward-deployed and embedded-engineering models now proliferating actually change delivery outcomes or simply repackage existing services, a distinction that will become apparent as the earliest engagements reach their second and third use cases. The second is how quickly agentic workloads move from the pilot column to the production column, since they concentrate all the difficulties described here: more governance surface, more integration points, and considerably more tokens per outcome. Organizations that have solved the handover problem at scale before agents arrive will find the transition a good deal cheaper than those that have not.

5 Lewis, M., & Shrout, R. (2026, May). The economics of agentic AI: On-premises deployments with Dell AI Factory with NVIDIA vs. cloud [White paper commissioned by Dell Technologies and NVIDIA]. Signal65. https://signal65.com/research/ai/the-economics-of-agentic-ai/

Information About This Report

AUTHORS

Nick Patience
Vice President & Practice Lead,
AI Platforms | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DELL Technologies

ABOUT DELL TECHNOLOGIES

Dell Technologies is a global technology leader that helps enterprises design, deploy, and operate modern IT infrastructure across data center, cloud, and edge environments. Through its Dell AI Factory approach and comprehensive AI implementation services, Dell enables organizations to accelerate the transition from AI experimentation to scalable production. The AI Factory brings together optimized infrastructure, software ecosystems, validated solutions, and expert professional services to deliver repeatable business outcomes. By embedding AI capabilities directly where data resides, Dell lowers latency and enhances security while supporting diverse workloads. Backed by strategic partnerships and robust delivery models, Dell equips enterprises to operationalize AI with greater reliability, performance, and cost efficiency.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum®
CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2026 The Futurum Group. All rights reserved.