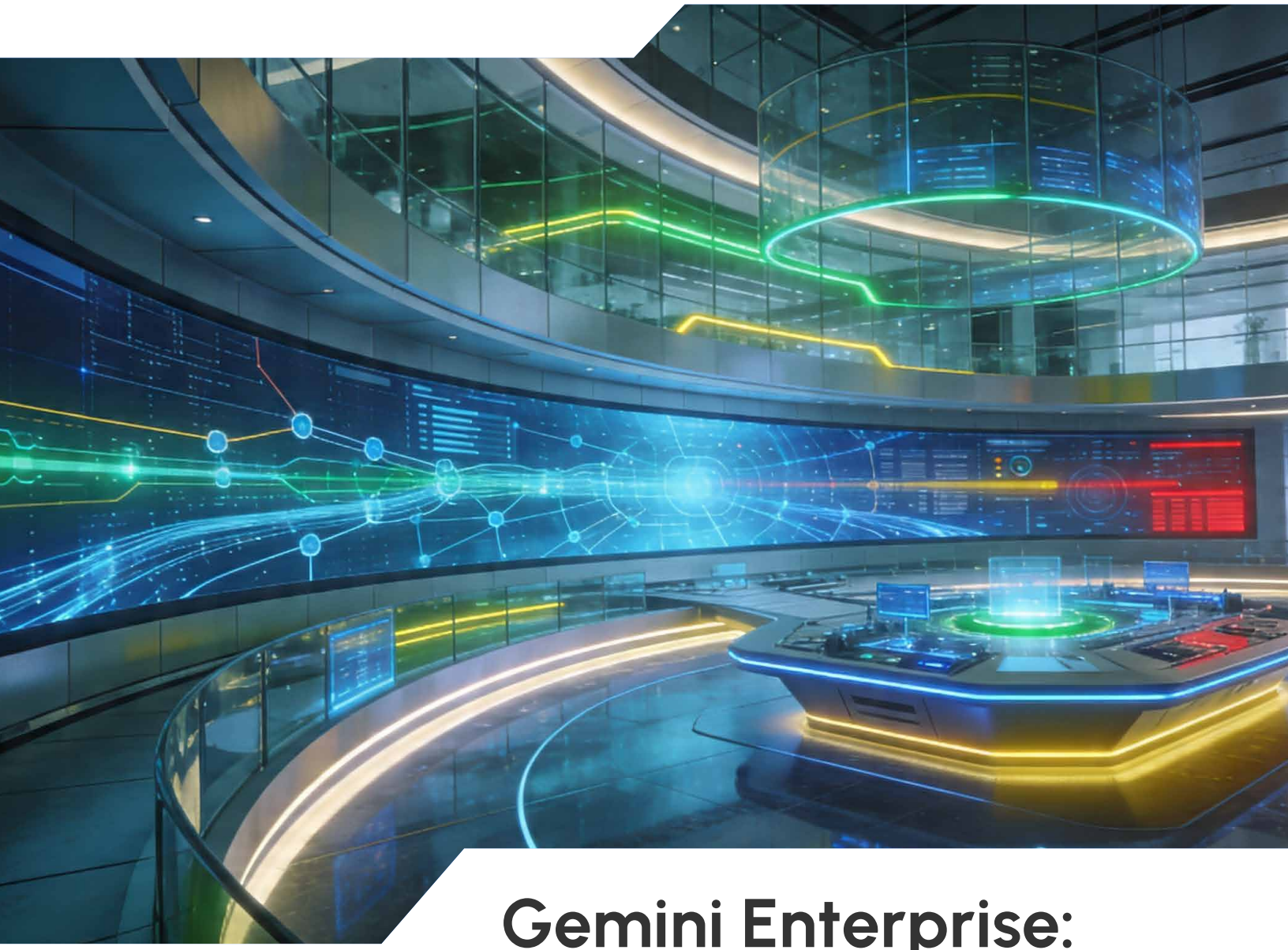
Futurum®

Google Cloud



Gemini Enterprise: Governing and Scaling the Agentic Enterprise

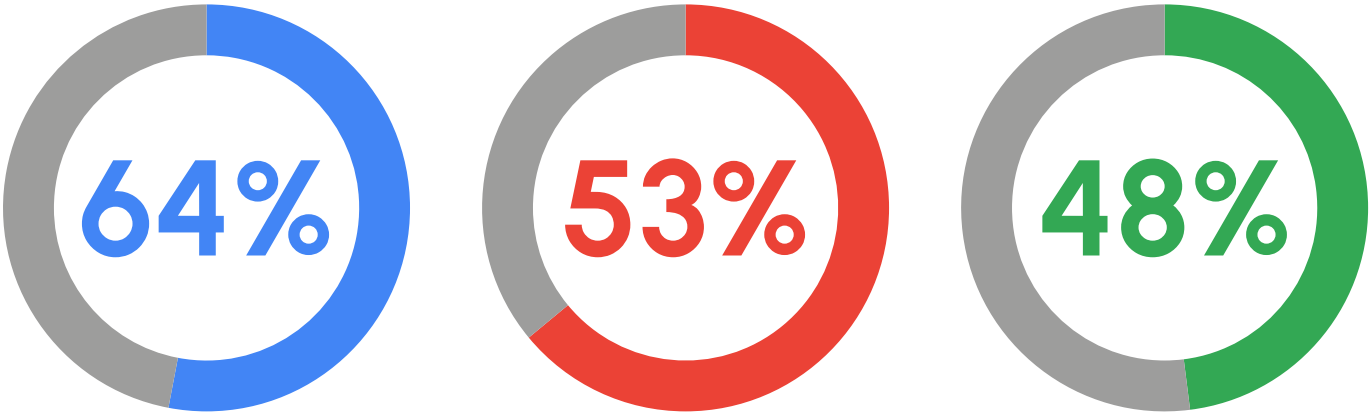
APRIL 2026

Introduction: The Shift from Pilot to Production

Enterprise adoption of agentic AI has moved from early experimentation toward operational deployment. Organizations are no longer evaluating whether to build AI agents but working through how to deploy them at scale, across functions, and within the governance frameworks that complex environments require. The 1H 2026 AI Decision Maker Survey (n=838), conducted by Futurum Research, found that 64% of organizations have moved beyond initial research and are now actively piloting or deploying agentic AI solutions, with customer service (52.8%) and IT operations (48.0%) as the leading areas of focus (see Figure 1).

The shift from pilot to production is not simply a matter of scale. Pilots typically involve a small number of agents, defined tasks, and close human oversight. Production deployments require agents to operate across systems and over extended time horizons, with varying degrees of autonomy. The infrastructure requirements are different in kind, not just in scope. The platform question is therefore not simply which AI model to use, but whether the surrounding infrastructure can manage agent identity, access, observability, and compliance at the level enterprise operations demand.

Figure 1. Adoption and Primary Use Cases for Agentic AI Among Organizations



Source: 2026 Futurum Group AI Decision Makers Survey, n=838

This move from pilot to production is happening alongside a massive rebalancing of enterprise priorities. According to Futurum Research's 1H 2026 Data Intelligence, Analytics, and Infrastructure Decision Maker Survey, data teams are pivoting hard from aspirational goals to operational execution, with measurable outcomes such as 'new business opportunities' and 'SLA attainment' taking precedence. However, they are hitting a human bottleneck: Acute skills shortages have more than doubled to 10.4%, replacing budget as the critical constraint.

This is where the underlying infrastructure becomes the differentiator. Organizations cannot simply hire their way out of this complexity. They require a unified data estate, as with the tight integration between AI and data seen with Google Gemini in BigQuery. Such moves abstract away the operational complexity, which 12% of our surveyed leaders still cite as the number one factor causing AI project failures.

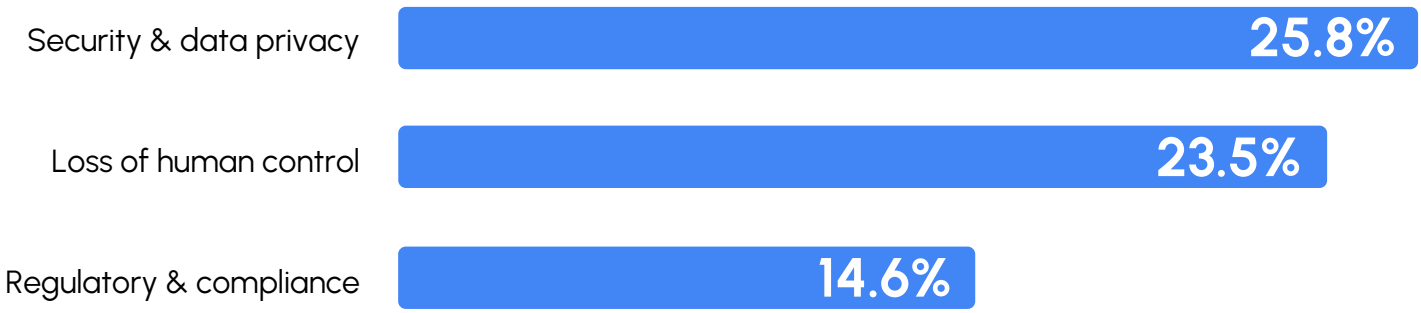
The Barriers to Agentic Scale

The central challenge for enterprises scaling agentic AI is not model capability. It is the absence of management infrastructure capable of governing an active agent population at production scale. Futurum research identifies three barriers that consistently prevent organizations from moving beyond the pilot stage.

Governance and Visibility

As the number of agents in an environment grows, so does the operational risk associated with incomplete visibility into what those agents are doing, what systems they can access, and whether their behavior conforms to policy. Without a centralized catalog of deployed agents and a mechanism to attach and enforce policies, IT and security teams face an expanding surface of autonomous activity that is difficult to audit or control. Futurum survey data indicates that security and data privacy vulnerabilities (25.8%) and loss of human control over critical decisions (23.5%) are the top concerns enterprise leaders associate with agentic AI deployment; both items are addressed with Gemini Enterprise (see Figure 2).

Figure 2. Top Enterprise Concerns with Agentic AI Deployment



Source: Futurum Group AI Decision Makers Survey, n=838

Integration and Context

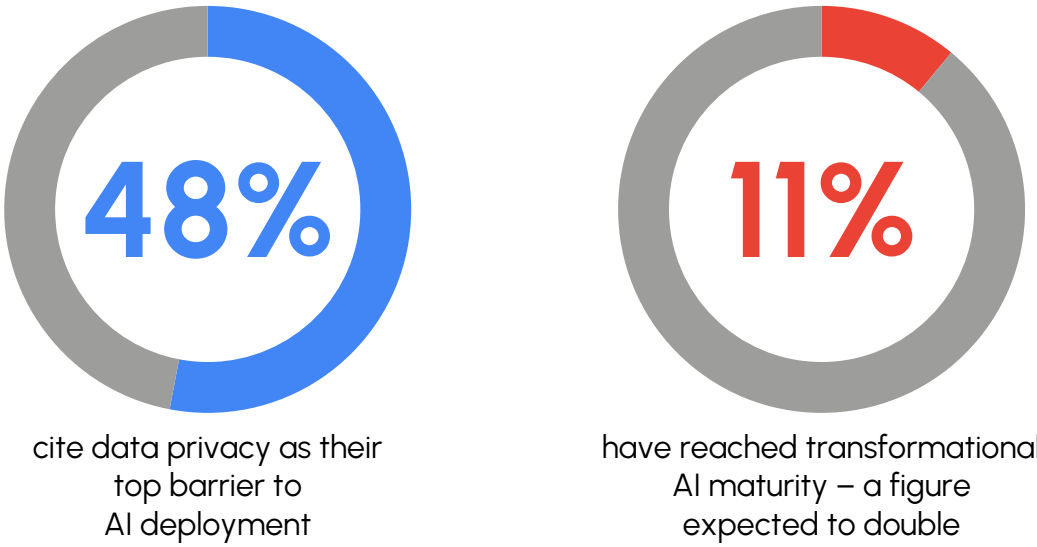
Agents limited to a single session or a single system can demonstrate value but cannot automate the multi-step, cross-application workflows that justify production investment. Persistent memory and cross-boundary operation are prerequisites for the kind of end-to-end automation that makes the business case for agentic AI at scale. Furthermore, the market is currently hitting a 'Write-Back Wall'. While reading massive amounts of data is a solved problem for agentic solutions, Futurum survey data reveals that the top two architectural bottlenecks for autonomous agents are integration complexity (29.3%) and the lack of transactional/OLTP write-back capabilities (24.6%).

An agentic platform cannot fulfill its ROI promises if it cannot safely execute transactions within systems of record. A production-grade platform must seamlessly bridge the analytical and operational divides. For companies deeply steeped in the Google Cloud ecosystem, by natively tying the Gemini Enterprise Stack into Google Cloud's operational databases (such as Spanner and AlloyDB) and analytical engines (BigQuery), enterprises can give agents the secure, governed authority to not just summarize data, but act on it. However, with Gemini Enterprise, Google can extend this same capability to customers wherever they are, on-premises, cross-cloud, or SaaS, making Gemini an excellent choice for companies seeking both stability and optionality in building out agentic workflows.

Observability and Evaluation

Deploying agents into consequential workflows requires confidence in how they will behave. The ability to simulate runs, evaluate outputs, trace reasoning, and improve behavior iteratively is what converts a successful pilot into a trustworthy production system, and its absence is a consistent reason pilots stall. This challenge is reflected in broader enterprise sentiment (see Figure 3): While nearly half of organizations cite data privacy as their top barrier to AI deployment, only a small minority have reached transformational AI maturity.

Figure 3. Barriers to AI Deployment and Current Maturity Levels (n=838)



Source: Futurum Group AI Decision Makers Survey, n=838

Deploying agents into consequential workflows requires more than just model evaluation; it requires a rigid, deterministic definition of business truth. If an agent and a human CFO define 'revenue' differently, the system fails.

This is why the Semantic Layer has emerged as mission-critical infrastructure for building trust in AI. Futurum's 1H 2026 market sizing forecast projects Semantic Layer investments to grow at a massive 19% CAGR, with 44.5% of enterprises planning to increase spend in this area to combat AI hallucinations (the top GenAI reservation for 24.9% of buyers).

The combination of Gemini Enterprise with Google Looker's semantic modeling ensures that agents are grounded in governed, universal business metrics. For example, enterprise practitioners building in Agent Designer and Agent Studio can readily bring in structured, governed data to better bridge the gap between raw data and an agent's reasoning engine. This empowers employees to shift away from manual coding tasks toward what Futurum terms as AI Shepherds. In fact, 72.5% of data professionals already rate themselves as highly focused on AI validation and storytelling over manual SQL coding, highlighting the critical need for platforms that prioritize observability and semantic trust over raw pipeline building .

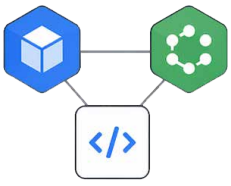
What a Production-Grade Agentic Platform Requires

The barriers above define the capabilities a production-grade agentic platform must address. Based on Futurum Research and enterprise engagement, five capabilities are foundational for organizations building a scalable, governed agentic infrastructure.

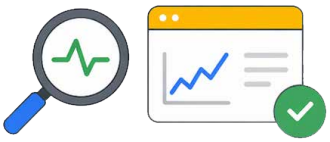
A unified governance and identity layer. Each agent operating within an enterprise environment should carry a defined identity, enabling the access controls that apply to human users to be consistently extended to agents. A centralized registry of deployed agents, with the ability to attach and enforce policies, is the basis for responsible agentic operations at scale.

Persistent memory and long-context support. Enterprise workflows rarely complete in a single session. Agents need to access information from the entire enterprise data estate while also retaining context across interactions and over extended time periods to support the end-to-end workflow automation that generates the most measurable operational value.

Broad model and integration flexibility. Organizations run inference across multiple model providers and operate across a mix of application environments. An effective agentic platform supports deployment across first-party and third-party models and integrates with existing enterprise systems through a range of connection patterns, without creating new dependencies that constrain future choice.



A coherent path from no-code to full-code development. Agentic AI deployment involves both technical and non-technical participants. Platforms that provide a continuous path from visual, no-code agent design to full-code development, without requiring agents to be rebuilt as they move between environments, reduce the time and friction involved in taking agents from initial design to production.



Observability, simulation, and evaluation. The ability to evaluate agent behavior before production, trace reasoning during operation, and improve performance iteratively is what makes governance commitments operational rather than aspirational. Without it, the confidence required to deploy agents into consequential workflows is difficult to establish.

Gemini Enterprise

Gemini Enterprise addresses these requirements through three connected surfaces aimed at knowledge workers, developers, and end customers: The Gemini Enterprise Agent Platform for building, scaling, governing, and optimizing enterprise ready agents; the Gemini Enterprise App serving as the front door to AI in the workplace; and Gemini Enterprise for Customer Experience for the customer journey – each serving a distinct set of users and workflows while running on shared governance infrastructure and a common set of controls. The controls are provided as a standard feature, eliminating extra cost and supporting organizations in their efforts to adopt agents at scale.

Gemini Enterprise Agent Platform: Build, Scale, Govern, Optimize

The Agent Platform is the foundational layer of the Gemini Enterprise stack, organized around four core areas that map to the lifecycle of a production agent deployment.

Build encompasses the Agent Development Kit (ADK), the Agent Studio, which provides a continuous path from no-code visual design to full-code development within a single environment – support for more than 200 first- and third-party models including Gemini and Claude through the Model Garden, and native support for agentic interoperability standards, such as the Agent2Agent Protocol (A2A), Model Context Protocol (MCP), Agent Payments Protocol, Universal Commerce Protocol (UCP), and Agent-to-User Interface (A2UI). The Agent Garden provides a library of pre-built and community-contributed agents as a further accelerant to deployment.

Scale capabilities include the Agent Runtime, Agent Sessions, Agent Sandbox, and the Agent Memory Bank. The Memory Bank provides persistent, long-context memory for agents, enabling workflows that operate across extended time periods without losing state. This directly addresses the constraint that has limited many enterprise agentic deployments to short-horizon, bounded tasks and is foundational to the kind of multi-step, cross-system workflow automation that justifies production investment.

Govern provides the identity, policy, and oversight infrastructure that enterprise production deployments require. Agent Identity gives each agent its own identity within the organization's access management framework, extending role-based access controls from human users to agents. Agent Registry provides a centralized catalog of all agents running within the environment, with the ability to attach and enforce policies at the agent level. Agent Gateway coordinates agent activity across applications and platforms, providing a governed path for agents to operate across system boundaries. Agent Anomaly Detection, Agent Security, and Agent Compliance round out the governance layer.

Gemini Enterprise's Agent Identity and Registry, when coupled with Google Cloud's broader data governance framework, can reduce operations complexity and cognitive overhead for those tasked with managing solutions in production. Whether building within a private cloud or by unifying Google Cloud foundational storage or database services such as Google Cloud Storage or BigQuery, companies can build out agentic governance within a single security perimeter, collapsing the need to govern and control access at multiple levels.

Optimize brings Agent Evaluation, Agent Simulation, Agent Observability, and Agent Optimizer. The simulation and evaluation capabilities allow organizations to assess how agents will behave across a range of real-world scenarios before deployment, observe performance in production, trace reasoning when outcomes are unexpected, and improve agent behavior in an iterative and evidence-based way. This closes the loop between agent deployment and agent governance, making the Govern pillar's commitments operationally meaningful.

Gemini Enterprise App: The Employee Experience

The Gemini Enterprise app serves as the AI interface for employees, designed to support the range of tasks that make up knowledge work, from individual research and content development to collaborative team workflows.

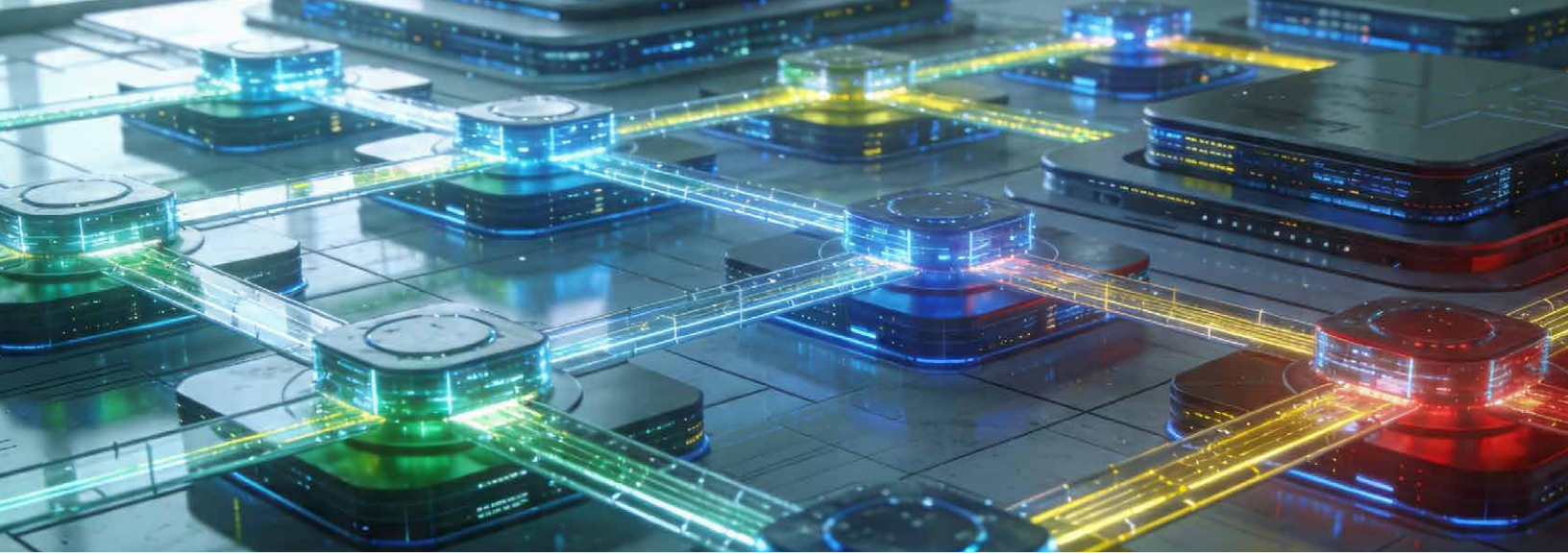
Canvas provides an interactive, co-editable document environment that allows users to develop and refine content collaboratively within the same interface, and to publish in both Google Workspace and Microsoft 365 formats, supporting mixed productivity environments. Projects provide shared, persistent context for teams: a contained instance of the Gemini Enterprise environment that retains full context as team members join or return, reducing the need to re-establish shared understanding at the start of each working session.

Agent Skills enable organizations to define repeatable, reusable agent capabilities that can be deployed and shared across the enterprise, reducing duplication and ensuring consistency. Inbox Alerts provide asynchronous notifications for long-running agents, decoupling the user from the need to actively monitor agent progress. The mobile app extends Gemini Enterprise to frontline and mobile workers.

Unified Context combines personal context drawn from an individual's working environment with enterprise-wide context from the organization's data and knowledge graph, so that agents operate with awareness of both individual working patterns and organizational knowledge simultaneously. Bring-Your-Own MCP (BYO MCP) support allows organizations to connect existing MCP-compatible tool endpoints directly into the environment, giving enterprises more control over their integration architecture and reducing dependence on pre-built connectors.

Gemini Enterprise for Customer Experience

Gemini Enterprise for Customer Experience extends the agentic platform to customer-facing workflows, providing a consistent customer context across touchpoints, from discovery and purchase through to service, so that customers do not need to re-establish who they are or what they need each time they engage through a different channel. This continuity, delivered through an omnichannel gateway, is the foundational requirement for agentic customer experience to function as a coherent journey rather than a set of isolated interactions.



Futurum Perspective

The governance capabilities in the Agent Platform – Agent Identity, Agent Registry, Agent Gateway, and the Optimize pillar – map directly to the barriers Futurum Research identifies as the primary obstacles to scaling agentic deployments. The 26% of enterprise leaders who cite security and data privacy as their top concern, and the 24% who cite loss of human control, are describing an infrastructure gap. Agent Identity extends existing access management frameworks to agents; Agent Registry provides the centralized visibility that policy enforcement requires; and the simulation and evaluation capabilities in the Optimize pillar give organizations a practical basis for moving agents into production with confidence.

The Agent Memory Bank and multi-model support address the other dimension of production readiness. Futurum's data shows that 71% of organizations use cloud APIs for inference and 51% run inference on their own infrastructure, often simultaneously; a pattern that makes broad model support a practical requirement rather than a differentiator . The Memory Bank enables the long-running, context-persistent agents that multi-step workflow automation requires, which represent the deployment pattern most likely to generate measurable enterprise value. Together, these capabilities position Gemini Enterprise to support organizations at their most critical stage as they move from demonstrated pilot value to governed, scalable production deployment.

Conclusion

The agentic enterprise is not a future state; it is a current deployment challenge. Organizations that have demonstrated the value of AI agents in pilots are now confronting the governance, integration, and observability requirements that scaling those deployments into production demands. Meeting those requirements calls for platform infrastructure that goes beyond model access to address the full lifecycle of agent deployment and governance: security, identity, policy, memory, observability, and evaluation.

For organizations building or refining their enterprise AI platform strategy, Gemini Enterprise offers a coherent answer to the production-readiness question – one grounded in the governance and lifecycle management capabilities that the transition from pilot to scale requires.

Important Information About This Report

AUTHORS

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

Brad Shimmin
Vice President & Practice Lead, Data Intelligence,
Analytics, & Infrastructure | 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

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.

Google Cloud

ABOUT GOOGLE CLOUD

Google Cloud delivers a comprehensive platform for building, managing, and scaling modern digital infrastructure across industries. Its suite of cloud services—spanning compute, data, AI, and networking—helps organizations accelerate innovation while maintaining high reliability and security. With the emergence of agentic AI, Google Cloud is enabling systems that can reason, act, and adapt autonomously, extending beyond traditional automation to deliver context-aware assistance and intelligent orchestration across environments. This agentic approach enhances efficiency and predictability for developers and enterprises alike, ensuring consistent, trustworthy performance in dynamic workloads. Together, these capabilities position Google Cloud as a leader in the next generation of intelligent, resilient computing.

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**