Deloitte. | Google Cloud

May 2026

SCALING AGENTIC AI
TO REALIZE BUSINESS VALUE

Prepared by Deloitte AI Institute in collaboration with Google Cloud

CONTENTS

PREFACE

The current wave of Artificial Intelligence (AI) impacting enterprises around the world is at a pivotal point. For the past three years, organizations have been experimenting with AI extensively. Now, they are looking to scale these initiatives and ***realize tangible business value.***

However, many are running into multiple hurdles along the way. These hurdles represent more than just technology concerns. In fact, they represent a complex intersection of strategy, organizational culture, and technical infrastructure. To scale successfully and realize value, enterprises would be well served to adopt a unified approach that combines the following:

Business and organizational

- Tightly align overarching business strategy with value generation.
- Define "agentic-native" business processes that reimagine, rather than replicate, existing workflows.
- Drive adoption and stickiness with sophisticated and sustained organizational change management.

Technical

- Establish ongoing platform operations and robust, scalable architectures.
- Integrate rigorous security and compliance practices.
- Adopt advanced engineering techniques suitable for learning and autonomous systems.

Governance

- Implement centralized governance frameworks that oversee both business outcomes (investments, initiative prioritization, execution rigor, measurement of value realization).
- Embrace trustworthy use of AI, including guardrail definition, security, and responsible use.

Bringing these elements together will be important for success. Below are several actionable steps organizations can take to successfully scale Agentic AI.

Our purpose

This paper, the third in a series on applying Agentic AI for business process reimagination, offers a clear, broad framework for scaling Agentic AI, with a focus on technical approaches and architecture, emphasizing real-world applications and actionable strategies.

This paper explores the transformative work at a major technology solutions firm and at a prominent insurance group Additionally, the report showcases the powerful synergy between Google Cloud's technology stack (including Gemini Enterprise) and Deloitte's significant industry and domain experience.

01

THE PIVOT POINT: *MOVING FROM EXPERIMENTATION TO THE AGENTIC ERA*

Background: the current state of enterprise AI adoption

For the last several years, the enterprise technology landscape has been dominated by the rapid democratization of Generative AI (GenAI). Organizations quickly moved from asking "what is this technology?" to "how can we use this?" This enthusiasm launched thousands of pilots and proofs-of-concept (PoCs) across industries. Examples include the rise of chatbots capable of summarizing dense email threads, marketing assistants that could draft copy quickly, and coding assistants that accelerated software development cycles.

However, an important realization has now taken hold in the boardroom: experimentation is not transformation. While 84% of organizations are increasing their AI investments, and 78% of leaders report greater confidence in the technology, productivity gains have often been limited to isolated pockets.[1] The massive, bottom-line impact promised by AI remains elusive at enterprise scale, keeping many companies at the edge of large-scale AI-driven transformation. This disconnect has created a "scaling wall," evidenced by the fact that only 25% of organizations have successfully moved 40% or more of their AI experiments into production to date.[2] Companies may now struggle to move from isolated GenAI pilots to deployed, governed, and integrated systems that fundamentally reshape how business is done.

The hurdle is no longer simply about access to powerful models; it is about the architecture of work itself. Traditional GenAI models are reactive. They wait passively for a human prompt to generate text or images. They are powerful, but they require constant human intervention. The pivot point we face today is the transition from these passive tools to Agentic AI: systems that can reason, plan, and execute complex workflows autonomously.

The scaling gap

The chasm between a successful pilot and a fully scaled agentic enterprise is defined by three primary friction points that leaders should address:

1. Technical fragility: Early pilots often lack the rigorous engineering discipline required for enterprise-grade reliability. A prototype that works well for five users may crumble under the weight of five thousand. Latency, cost management, and error handling that is manageable in a sandbox become important failures in production.

2. Process inertia: Organizations often tend to apply AI to existing, inefficient processes, effectively "paving the cow path" rather than reimagining the journey. True transformation requires deconstructing workflows to their atomic elements and rebuilding them as "agentic-native" processes, rather than simply adding an AI layer to a legacy workflow.

3. Trust deficits: Without robust observability and governance, business leaders remain hesitant to let AI "act" on behalf of the company. The fear of "hallucinations" (fabricating facts) or "runaway agents" (executing unintended actions) creates a paralysis where powerful tools are restricted to low-impact tasks. Bridging this trust gap requires new mechanisms for visibility and control.

The rise of Pragmatic AI: from hype to ROI

To overcome the scaling wall, organizations should adopt a **Pragmatic AI** mindset that prioritizes functional results over technical novelty. This approach represents a deliberate shift away from the industry hype and toward a disciplined, value-first strategy. This mindset is defined by three interconnected shifts that help transform AI from a laboratory experiment into a reliable business asset:

Value over velocity: Organizations should move beyond the race to build bots quickly and instead focus on the specific bottom-line value each agent creates.

Specialized utility over Artificial General Intelligence (AGI): Rather than chasing general intelligence, leaders should prioritize specialized utility. This involves designing agents to execute high-impact processes—such as invoice matching or contract review—with hyper-reliability.

Outcome-driven roadmaps: Experimental curiosity should give way to rigorous assessment. AI investments should follow a roadmap dictated by projected ROI and measurable business impact.

02

DEFINING THE AGENTIC SHIFT: *THE COGNITIVE LEAP*

Beyond traditional AI and GenAI

To scale effectively, it is important to first define precisely what is scaling. Agentic AI represents a fundamental evolution from the "stochastic parrots" of early GenAI. Where traditional (discriminative) AI focused on classifying data and making predictions based on historical patterns, and Generative (creative) AI focused on creating new content based on learned patterns, **Agentic (active) AI** is focused on achieving goals. It moves the technology from a tool that talks to a tool that *does.*

The core characteristics of agents

As defined in the research, an AI agent is not merely a language model; it is a sophisticated **reasoning engine.** It possesses a cognitive architecture that allows it to operate with a degree of independence previously unattainable.

This architecture is built on six pillars:

Perception: The ability to understand context from multimodal inputs, processing text, images, and audio simultaneously to form a complete picture of the user's intent.

Planning: The capacity to break down a high-level, ambiguous goal into a sequence of logical, executable steps without explicit human instruction for each micro-task. For example, "resolve this customer dispute regarding the Q3 invoice."

Tool Use: The ability to connect to and manipulate external systems, such as ERPs, CRMs, APIs, and web browsers, to fetch live data or perform actions in the real world.

Action: The execution of a plan, navigating systems and interfaces to complete the work.

Memory and Reflection: The retention of context over time, to help enable the agent to learn from past interactions and improve its performance, creating a personalized and increasingly effective experience.

Orchestration: The ability to collaborate with other agents to perform a larger task or an entire workflow by each agent taking on a task specific to its own area of speciality and integrating upstream and downstream with other agents.

The power of multi-agent systems

While a single agent is powerful, the true enterprise unlock lies in **multi-agent systems (MAS)**. In a MAS, the workflow is not handled by a generalist model, but by a team of specialized agents that collaborate much like a human team.

Imagine a complex procurement process. In a MAS architecture, one agent might act as the "legal reviewer," trained specifically on contract law and compliance. Another agent might serve as the "data extractor," specialized in pulling data from unstructured PDFs. A third agent acts as the "negotiator," handling vendor communications. By decomposing complex processes into these specialized roles, enterprises can achieve higher accuracy, as each agent stays within its domain of knowledge. This modularity also helps enable easier debugging and governance, as the failure of one component does not collapse the entire system.

While an agent's ability to think and act is impressive, these 'cognitive' skills don't mean much if they can't work within a professional business environment. Moving from a single smart demo to a network of reliable AI assistants requires more than just clever instructions; it requires a real foundation. To make these agents work safely and at scale, we need a digital home base—a secure system that provides the memory and data that agents need to be useful. In the next section, we'll look at the Agentic Operating System that turns these individual AI traits into a powerful business reality.

03

THE TECHNICAL FRAMEWORK FOR SCALE: *BRINGING THE ELEMENTS TOGETHER*

To move past the "scaling wall," organizations should adopt a holistic framework that integrates various key elements of a scalable AI architecture. This is not a few different pieces that are integrated together, but a unified system where each element reinforces the others.

Scaling requires a shift from simple "prompt engineering" to rigorous "system engineering." This involves building a foundation that helps ensure reliability, security, and access to knowledge.

Experience indicates that scaling requires the following elements to come together:

1. The developer workbench

A robust surface/platform for development that helps enable a wide variety of frameworks and runtime options to develop agents and multi-agent systems, as well as services such as state management, memory, vector search etc. This is also the surface from which the developer can consume various services provided for agentic operations.

2. An agentic operating system

A core system comprising multiple services to help enable effective management of the full agentic life cycle, including:

- a. Common services: reusable frameworks and templates, tools and services that help enable developer productivity, a unified gateway for various tools and registries, Continuous Integration / Continuous Delivery (CI/CD)[EG1] tools.
- b. AgenticOps: operational capabilities such as observability and evaluations that help enable monitoring, measuring and enhancing the performance, reliability, and quality of the agentic system.
- c. FinOps: the ability to set, monitor and enforce thresholds related to consumption of various key services that directly impact the economics of running an agentic system.
- d. Governance and security: automatic discovery of agents and managing agent identity and permissioning, and defining and enforcing guardrails at runtime[EG2].

3. A model and agent repository

Access to a wide variety of off-the-shelf and custom models as well as existing agents to help enable the use of tested and curated models, while also enabling reuse of existing and proven agents.

4. Enterprise context:

A knowledge layer that helps enable building the enterprise context by creating a long term context graph that captures agent decision traces and combines it with enterprise data to provide richer context during runtime.

5. Data foundation:

A scalable data foundation that brings important enterprise data in a consumable manner.

6. Cloud foundation:

A core set of cloud services that help enable a scalable environment to host and deliver enterprise-scale agentic systems.

In an agentic world, governance cannot be a bottleneck that stifles innovation; it should be a real-time capability that helps enable safe speed.

Achieving agent observability (AgentOps)

It’s not possible to manage what cannot be seen. Enterprises need Agent Operations (AgentOps), a dedicated layer of the technical stack that monitors agent performance and behavior across the organization. In a robust enterprise architecture, this is powered by the AgentOS Operations Layer, a central control plane within a centralized hub that bundles observability, shared services, and core governance to manage the life cycle and compliance of agent fleets.

Accessible through a dedicated operations interface (part of the dual-faceted AgentOS interface layer), developers as well as site reliability engineers (SREs) gain a unified, human-in-the-loop entry point to monitor agents deployed across various project-specific "spokes." This observability stack aggregates data via various tools, operating on two distinct paths:

• **Real-time monitoring and intervention:** Monitors active workflows for immediate risks, such as toxic outputs, personally identifiable information (PII) leakage, or regulatory violations. By integrating with a centralized policy enforcer, the hub can trigger immediate circuit breakers to halt an agent's execution across any spoke if non-compliant behavior is detected. Traceability tools help operators to pinpoint exactly where an agent's reasoning diverged in real time.

• **Over-time aggregate analysis:** Analyzes long-term agent performance using aggregated metrics from interconnected spokes. This includes using specialized model monitoring to detect "drift" in reasoning quality or model degradation. Additionally, a dedicated FinOps module tracks the cost per transaction, while pipeline monitors track latency and efficiency trends. This telemetry informs continuous, long-term optimization, feeding directly into model tuning services to iteratively refine underlying models and agent accuracy based on real-world enterprise interactions.

To mitigate "agent sprawl," a scenario where thousands of unvetted, duplicative, and costly agents proliferate across isolated company silos, organizations should implement internal AI agent marketplaces powered by the hub’s custom services. These marketplaces act as centralized, shared repositories, using a specific agent registry and a prompt registry, where vetted, compliant, and secure agents are published.

Here is how this ecosystem helps ensure control:

Standardized development (spokes): developers build agents in their project-specific AgentOS developer layer (spoke) using standardized frameworks (such as ADK, LangGraph, or CrewAI).

They use shared enterprise tools like Context Caching and the Vertex AI Memory Bank for state management. This work is done through the Developer Interface.

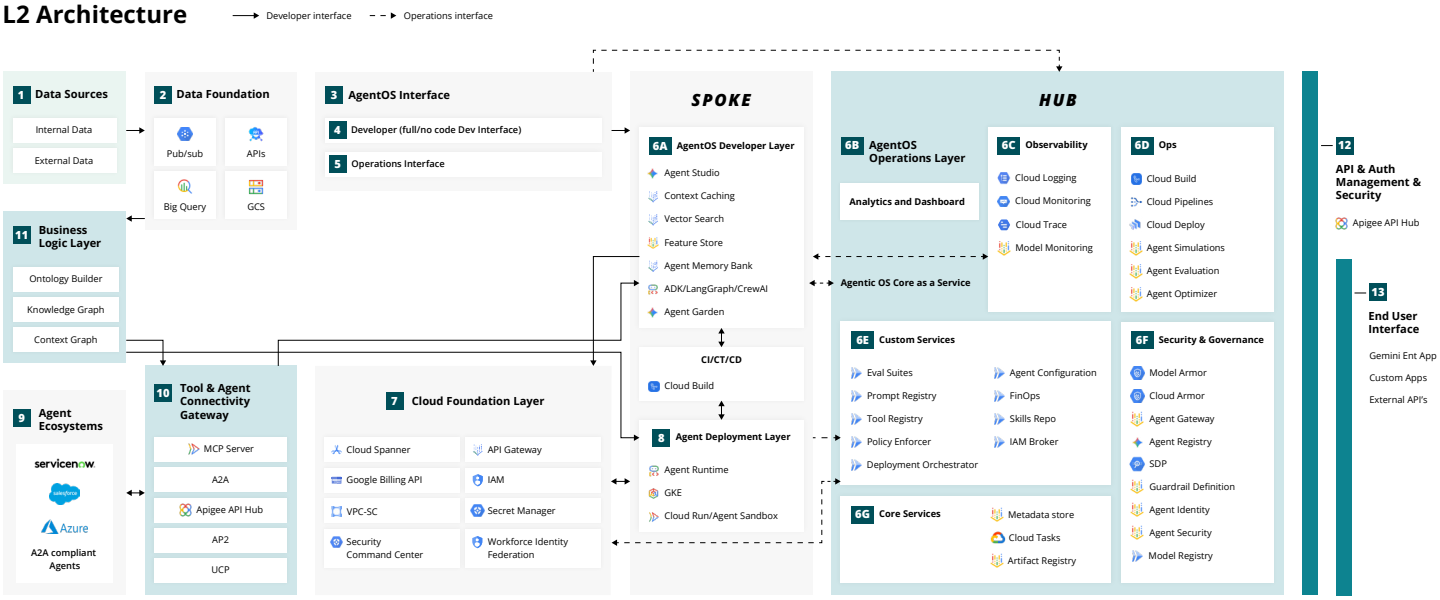
Governed deployment and secure execution (hub): once an agent is built, it doesn't just go live blindly. It moves through the Ops (hub), using Cloud Deploy for standardized, policy-checked rollouts. Furthermore, for agents dynamically generating and executing scripts, the platform uses a secure Agent Sandbox; an isolated runtime environment that supports agents as they execute code safely, without exposing core systems to vulnerabilities or unintended destructive actions.

FinOps as value driver: in the "agentic operating system," the FinOps module is not just a monitoring tool; it is the engine of Pragmatic AI. By enforcing consumption thresholds and tracking the exact cost per transaction, it ensures that the cost of an agentic workflow rarely eclipses the value it generates.

Reuse via marketplaces: the AI Agent Marketplace reduces "agent sprawl" and redundant development costs. By promoting the reuse of vetted, standardized agents (for example, a single "Invoice Processing Agent" for global regions), organizations realize a faster return on their initial engineering investment.

By centralizing operations, governance, security, and registries into a foundational hub, the enterprise helps to ensure consistency, reduces redundant compute costs (via Context Caching and FinOps), and promotes adherence to corporate policy while enabling individual business units to scale efficiently.

Below is a typical AgentOS architecture, using natively available services on the Google Cloud Platform.



Components of Agent OS

DATA LAYER

1 & 2. Data layer: The data processing backbone responsible for the Ingestion (Pub/Sub), Transformation, and Target storage (BigQuery, Databricks) of structured and unstructured data used by agents consisting of Data Foundation and Data Sources.

3, 4 & 5. AgentOS interface layer: The dual-faceted portal system comprising of the developer interface for building agents and the operations interface for managing the platform, providing the primary human entry points into the system.

6A. AgentOS developer layer (spoke): Project-specific build environment where developers author agents using Agent Studio and ADK/LangGraph/CrewAI. Includes Vector Search, Feature Store, and Agent Memory Bank for grounding and state, Context Caching for performance, and an in-spoke Agent Garden for reusable templates.

6B. AgentOS operations layer (hub):
The central control plane bundling observability (logging, monitoring), shared services (registries, policy enforcement), and core services (security, data governance) to manage the lifecycle and compliance of agent fleets.

6C. Observability (hub): The central monitoring stack within the hub that aggregates logs, traces, and metrics from spokes via cloud logging, cloud trace, and cloud monitoring, including specialized model monitoring for drift detection. Looker could enable real-time visualization of agent performance, costs (FinOps), and ROI metrics.

While infrastructure logs capture platform activity, agents must also be instrumented at the code level using OpenTelemetry (OTel). This means capturing distributed traces across each reasoning step and tool call, LLM-specific metrics such as token consumption, latency, and model version; and sanitized prompt/response logs streamed for quality auditing.

6D. Ops (hub): The centralized operations center for managing the automated lifecycle of agents. It utilizes Cloud Build and Cloud Deploy for CI/CD, Cloud Pipelines for orchestration, and includes specialized processes for Agent Simulations, Agent Evaluation, and performance tuning via Agent Optimizer.

6E. Custom services (hub): A collection of shared registries and governance tools: Prompt Registry, Tool Registry, Skills Repo, Policy Enforcer, Deployment Orchestrator, Eval Suites, modules for Agent Configuration, FinOps, and IAM Broker that standardize and control agent development across the platform.

6F. Security and governance (hub): This is a centralized suite of security controls within the Hub, responsible for protecting the platform and its agents. It includes services like Model Armor and Cloud Armor for threats, Agent Gateway, Model Registry and Agent Registry for controlled access, SDP for data privacy, Guardrail Definition, plus Agent Identity and Agent Security.

6G. Core services (hub): The foundational, shared services that underpin the AgentOS Operations Layer. This layer includes a central Metadata store for tracking artifacts and lineage, Cloud Tasks for managing asynchronous execution, and Artifact Registry for storing and managing build artifacts.

AGENT ECOSYSTEM LAYER

7. Cloud foundation layer: The underlying stack of GCP primitives that serves as the infrastructure bedrock for the operating system. It includes Cloud Spanner, Google Billing API, VPC-SC, API Gateway, IAM, Workforce Identity Federation, Security Command Center, and Secret Manager.

8. Agent deployment layer: The execution environment and CI/CD pipeline responsible for hosting agents on scalable runtimes. It now features Agent Runtime and Cloud Run/Agent Sandbox for containerized execution on serverless or GKE infrastructure, driven by Cloud Build pipelines.

9. Agent ecosystems: The external integration layer connecting the OS to third-party platforms (ServiceNow, Salesforce, Azure) and A2A-compliant external agents to extend capabilities beyond the core infrastructure.

10. Tool and agent connectivity gateway: The interoperability hub that utilizes standard protocols like the MCP Server, A2A, and UCP to facilitate secure communication between agents, external tools, and enterprise APIs via the Apigee API Hub.

11. Business logic layer: The semantic intelligence engine comprising Ontology Builders, Knowledge Graphs, and Context Graphs that grounds agents in enterprise-specific logic and relationships.

API AUTH AND INTERFACE LAYER

12. API, auth management and security: The protection layer (utilizing Apigee) that governs access control, secures API traffic, and manages authentication across the entire agent ecosystem.

13. End user interface layer: The consumption layer consisting of Custom Apps, External APIs, and Gemini Enterprise endpoints where business users interact with deployed agentic solutions.

The "agentic guardrail" architecture

Standard GenAI security often relies on static filters. For Agentic AI, security should be dynamic and multi-layered because agents have the power to act.

The four-tier defense model

To scale safely, enterprises should implement a "Defense in Depth" strategy for autonomous systems:

1. **Linguistic guardrails (input/output):** using models like Google's Shielded Gemma or specialized classifiers to detect prompt injection, "jailbreaking" attempts, and PII leakage in real-time.

2. **Behavioral guardrails (the sandbox):** as mentioned in the architecture, the Agent Sandbox is important. Agents shouldn't execute code or access databases in the open production environment. They should operate in "zero-trust" containers where their actions are limited by strict resource quotas.

3. **Semantic guardrails (the "constitution"):** implementing a "Constitutional AI" layer; a secondary, lightweight model that audits the primary agent's intent. Before an action is taken (for example, "Transfer US$5,000"), the Auditor Agent checks the plan against the company's "digital constitution" to help ensure policy compliance.

4. **Infrastructure guardrails:** while the above tiers secure the agent's logic, the underlying platform must also be hardened. Enterprises should leverage capabilities like Google's new Agentic Defense (integrated with Wiz). This unified security platform deploys specialized agents for machine-speed threat hunting and automated detection engineering, instantly identifying vulnerabilities across the entire AI infrastructure before they can be exploited by or through autonomous systems.

04

THE ROBUST AI PLATFORM: *CREATING SUPERIOR DEVELOPER AND USER EXPERIENCES*

Scaling Agentic AI requires a platform that solves the "plumbing" problems of infrastructure, security, and compliance, so that the business can focus on the value. A robust platform, such as **Google Cloud's Gemini Enterprise** and **Vertex AI**, helps to enable this by catering to two distinct but equally important experiences.

Superior developer experience: agentic orchestration

Developers are the architects of the agentic enterprise. They should not be bogged down building memory management systems or writing boilerplate API connectors from scratch. A robust platform provides a broad toolkit, consisting of:

- **Pre-built connectors:** Seamless integration with major enterprise systems (such as ServiceNow, Salesforce, Jira, Workday) out of the box, reducing integration time from weeks to minutes.
- **Orchestration layers:** Sophisticated tools to manage the hand-offs between specialized agents, to help ensure that context is preserved as a task moves from a "researcher agent" to a "writer agent" to a "compliance agent."
- **Vector search and grounding:** Integrated capabilities to "ground" agents in enterprise truth, reducing hallucinations by forcing the agent to cite its sources from internal data.

Identity and access management for agents (IDAM-A)

In a multi-agent system, an agent is effectively a digital employee. We propose building a robust framework for IDAM for agents.

Treating agents as identities

- **Machine-to-machine (M2M) auth:** each agent should be issued a unique digital identity, or workload identity. This helps enable security teams to track exactly which agent accessed which record in the ERP or CRM, providing an audit trail of thought.
- **Privilege scoping:** just as you wouldn't give an intern access to the entire HR database, an HR agent should have read access to specific tables. Organizations should move from full API access to granular scoping, where agents are granted the minimum viable permissions to complete a specific goal.
- **Attributable liability:** by assigning a human-in-the-loop (HITL) owner to each agent identity, the organization maintains a clear line of accountability for each autonomous decision.

Superior user experience: the "Intranet of agents"

For the employee, the complexity of the underlying models should be invisible. They need a unified, intuitive interface where they can interact with a "marketing agent" or an "HR agent" as naturally as they chat with a colleague.

1. **Discoverability:** users should be able to easily find approved agents relevant to their specific role or department.
2. **Accessibility:** agents should be integrated into the flow of work rather than hidden in a separate, disconnected portal. For example, the agents could reside directly within Google Workspace side panels.
3. **Trust and transparency:** users need to trust the tool. The interface should provide citations and direct links to source documents for each agent output, to help enable instant verification.

The agentic red-teaming life cycle

The agentic red-teaming life cycle shifts security from a static, one-and-done checklist into a continuous, rigorous loop, integrated across the entire development journey.

This begins in the development phase with adversarial simulation, where developers deliberately attempt to trick agents into hallucinating or seeking unauthorized access within the spoke/dev environment to harden their logic. Upon deployment, the system uses circuit breakers—automated triggers within the hub—that immediately revoke an agent's credentials if drift or anomalous behaviors, such as excessive API calls, are detected in real-time.

Finally, the post-action phase helps ensure accountability through forensic traceability, leveraging tools like Google Cloud Trace to reconstruct an agent's reasoning path after an error, which helps enable precise debugging and high-level auditability for enterprise trust. Many of these tasks could be automated by agents specialized in them.

05

THE HUMAN ARCHITECTURE: *BEYOND TECHNICAL IMPLEMENTATION*

The invisible costs of AI Transformation

The primary barrier to AI success is rarely the technology itself; rather, it is the invisible costs of organizational transformation. According to a recent study published by the Stanford Digital Economic Lab, approximately 77% of the challenges in enterprise AI implementation relate to change management, data quality, and process redesign.[3] Without a robust foundation, AI initiatives may remain confined to isolated, fragmented pilots that fail to deliver enterprise-wide value. To scale effectively, organizations should move beyond these silos and take an end-to-end business process reimagination approach.

Reimagining the workflow

A frequent enterprise pitfall involves treating AI as a pure technology project rather than a change management initiative. Applying AI to a broken workflow serves to amplify existing inefficiencies, making a bad process worse by making it faster. Successful adoption requires a deliberate cleanup phase—mapping workflows, identifying systemic bottlenecks, and simplifying redundant processes before agentic systems are deployed.

The role of executive sponsorship

Executive sponsorship should transcend passive budget approval. Realizing the promise of Pragmatic AI requires active steering where leaders engage consistently to clear organizational bottlenecks. Furthermore, successful rollouts are rarely tech-led in isolation. They require cross-functional co-sponsorship, pairing the CTO's technical vision with the CEO's strategic mandate and department heads who define the ultimate success metrics. Transformation is most profound when AI adoption is embedded directly into corporate objectives and key results (OKRs) and tied to performance incentives.

Fostering a culture of iterative failure

Because organizations rarely optimize agentic workflows on the first attempt, a culture that permits safe, iterative failure is non-negotiable. To help build and maintain this culture, leadership should actively sponsor continuity, while simultaneously building an atmosphere of psychological safety. This will send a clear message that experimentation is a strategic necessity, not a career risk. After all, if executive sponsors abandon ship after an initial setback, the organization loses the institutional memory of what failed and why. Instead, by keeping the scope of initial pilots controlled and treating them explicitly as experiments, companies can absorb failures, integrate user feedback, and eventually build highly successful, scalable systems.

Individual versus collaborative evolution

There is a stark divergence between individual and collaborative AI adoption. While GenAI has successfully shifted behaviors that workers manage independently—such as drafting documents or managing inboxes—it has yet to significantly alter systemic coordination, such as time spent in collaborative meetings. This highlights a fundamental reality: changing how one person works is a matter of tool adoption; changing how a team works is a matter of organizational transformation. To achieve systemic productivity gains, executives should actively redesign workflows and formally integrate AI into the company's collaborative culture.

06

REAL-WORLD APPLICATIONS: *AGENTS IN ACTION*

Leading organizations are moving beyond pilots to deploy agents that drive tangible business value in production environments.

CASE STUDY: A MAJOR TECHNOLOGY SOLUTIONS FIRM

The Challenge: a global leader in enterprise asset intelligence faced a significant operational bottleneck in their accounts payable (AP) workflows. The sheer volume of invoices and purchase orders (POs) arriving in diverse formats required extensive manual validation. This led to processing delays and diverted the skilled finance team from high-value analysis to rote data entry.

The agentic solution: In collaboration with Deloitte and Google Cloud, the organization deployed a sophisticated agentic workflow to automate the AP process. The solution did not just digitize the documents; it made the decision-making autonomous.

The Workflow:

- **Read and extract:** specialized agents used multimodal capabilities to ingest invoices and POs from PDFs, images, and emails, extracting key data fields with high accuracy regardless of the document layout.
- **Reason and validate:** the agents perform "3-way match" logic, cross-referencing line items between the invoice, the PO, and receiving data to identify discrepancies in price, quantity, or terms.
- **Flag and resolve:** crucially, the system manages exceptions. Perfect matches are processed for payment automatically (straight-through processing). When a discrepancy is found, the agent flags the specific anomaly and routes it to a human validator, providing the context needed for rapid resolution.

IMPACT

This shift transformed the finance function, streamlining data extraction and significantly reducing the manual workload. The finance team shifted their focus from data entry to exception handling and strategic financial management.

REAL-WORLD APPLICATIONS: *AGENTS IN ACTION*

CASE STUDY: A MAJOR EUROPEAN INSURANCE COMPANY

The Challenge: a major European insurance and financial services group possessed a vast repository of internal knowledge—HR policies, complex insurance tariffs, and procedural documentation. However, this knowledge was trapped in silos. Employees spent excessive time navigating SharePoint and Confluence to find basic policy answers, slowing down both internal HR operations and external customer service.

The agentic solution: the company implemented an AI knowledge assistant powered by Gemini Enterprise to unlock this trapped value.

The Workflow:

- **HR policy access:** agents ingest and index dense internal documentation. Employees can now ask natural language questions (e.g., "What is the policy for paternity leave in this region?"). The agent retrieves the specific clause, synthesizes a clear answer, and cites the source document.
- **Customer service support:** specialized agents assist human representatives by cross-referencing customer health data against complex insurance policies in real-time. The agent provides an immediate "coverage confirmation" or "policy exclusion" summary to the human rep.

IMPACT

This deployment transformed internal knowledge from a static asset into an active participant in business operations. It significantly reduced resolution times for employee inquiries and customer support tickets, leading to higher satisfaction and operational efficiency.

07

NAVIGATING THE MARKET: *OPPORTUNITY VERSUS NOISE*

The noise

The current market is flooded with "AI washing"—the rebranding of basic automation scripts or simple chatbots as "agents." There is a surplus of hype regarding "Artificial General Intelligence" (AGI) that often distracts leaders from the immediate, pragmatic value of "specialized utility." Leaders should be wary of demos that look impressive but lack the robustness to handle enterprise complexity.

The opportunity

The true market opportunity lies in **agentic process automation.** It is not about deploying a single AI that does everything; it is about deploying thousands of specialized, reliable agents that do specific things perfectly. The noise focuses on "cool demos"; the value focuses on "boring processes"—invoice matching, contract review, schedule coordination—executed at hyper-speed and scale.

08

STRATEGIC OUTLOOK: *THE KEY TAKEAWAYS*

Looking towards the "Agentic Enterprise 2028," several strategic imperatives emerge for leaders who wish to stay ahead of the curve.

Challenges to anticipate

- **Data gravity:** agents struggle if data is fragmented. The "dirty work" of data cleaning, consolidation, and vectorization is unavoidable and should be prioritized. Organizations can now bypass massive data migration efforts using Google's newly announced Cross-Cloud Lakehouse.

  Standardized on Apache Iceberg, this capability allows businesses to query data directly in AWS or Azure without moving it. This provides a frictionless foundation for agents to fetch context and act on data seamlessly, regardless of its original cloud location.

  To ensure agents truly understand this distributed data, organizations can utilize the Knowledge Catalog. It provides continuous data enrichment—going beyond manual curation to actively mine structured schemas, query logs, and BI semantic models while extracting entity relationships from unstructured data.
- **The "frozen middle":** middle management may resist agent adoption due to a perceived loss of control or relevance. Change management should specifically target this layer to turn managers into enablers rather than blockers.
- **Cost management:** without robust FinOps practices, "runaway agents" (agents that get stuck in loops or overuse expensive models) can drive up inference costs unexpectedly. Governance frameworks should include cost monitoring and budget caps.

STRATEGIC OUTLOOK:
THE KEY TAKEAWAYS

Strategies for success

- **Think "process," not "task":** don't just automate the writing of an email; automate the entire "customer outreach" workflow. Look for end-to-end value chains that can be reimagined.
- **Build the "digital org chart":** define roles for your agents just as you would for employees. Who is the "manager" agent? Who is the "doer"? Who is the "auditor"? Treat your digital workforce with the same structural rigor as your human workforce.
- **Invest in "Agent Ops":** establishing a command center for agent observability is non-negotiable for scaling trust. You should be able to inspect the "thought process" of your agents.
- **Adopt a "pragmatic first" filter:** each proposed agentic use case should be passed through a Pragmatic AI filter. Ask: Does this solve a documented bottleneck? Is the ROI measurable within 6-12 months? Can it be scaled without a linear increase in cost? This helps to ensure the "Digital Org Chart" is built on a foundation of economic viability.

Future outlook

A future of heterogeneous multi-agent ecosystems is emerging.

In the near term, or one to two years, internal "agent marketplaces" are anticipated to become standard within large enterprises. By 2028, cross-enterprise agent collaboration may emerge; for example, where a retailer's "inventory agent" negotiates directly with a supplier's "shipping agent" to restock goods autonomously, ***creating a frictionless, automated economy.***

09

THE POWER OF SYNERGY: *DELOITTE AND GOOGLE CLOUD*

The journey to an Agentic Enterprise is complex, but organizations do not have to walk it alone. The collaboration between Deloitte and Google Cloud offers an individually powerful engine for this transformation, combining leading technology with deep organizational wisdom.

Google Cloud's technology stack

Google Cloud provides the industry's only full-stack agentic capability, offering a secure foundation for the future:

- **Gemini family of models:** a multimodal AI model capable of complex reasoning and an industry-leading long-context window of up to 2 million tokens. This helps enable agents to "read" and reason over massive enterprise documents, such as legal contracts, codebases, or financial reports, without losing context.
- **Gemini Enterprise:** the secure, user-facing integration layer that embeds AI directly into the employee's daily flow of work. Acting as the intuitive "front door" to the agent ecosystem (accessible via Google Workspace side panels, Docs, and Gmail), it helps ensure strict enterprise data protections. Crucially, it helps ensure that proprietary company data, user prompts, and agent interactions are kept private and are not used to train Google's public models.
- **Vertex AI Agent Builder:** a developer-friendly platform that democratizes agent creation, and helps enable teams to build, deploy, and manage agents using natural language prompts rather than complex code.

Deloitte's industry and domain experience

Technology needs context to drive value. Deloitte brings the business and governance gears to the Google engine:

- **Industry advantage:** Deloitte offers pre-built agent blueprints tailored for specific sectors (for example, "Banking Onboarder" for the financial sector, or "Clinical Compass" for healthcare), accelerating time-to-value by solving industry-specific problems out of the box.
- **Trustworthy AI framework:** a proven methodology to help ensure agents are fair, transparent, and accountable, mitigating the risks of bias and hallucination.
- **Adoption services:** a human-centric approach to change management, using our "TrustID" framework to help ensure the workforce is psychologically and operationally ready to collaborate with their new digital colleagues.

The pivot to Agentic AI is not merely an IT upgrade; it is a fundamental expansion of the workforce and transformation of the business. By combining the robust, scalable infrastructure of Google Cloud with the strategic foresight and implementation rigor of Deloitte, enterprises can confidently scale their AI initiatives. This collaboration can turn the abstract promise of autonomy into the concrete reality of sustained business value.

Deloitte. | Google Cloud

AUTHORS

Gopal Srinivasan
Principal
Deloitte Consulting LLP
gosrinivasan@deloitte.com

Aman Azad
Global Alliance Leader for Deloitte Google Cloud
azadaman@google.com

Mamoun Hirzalla
Managing Director
Deloitte Consulting LLP
mhirzalla@deloitte.com

Ved Ramakrishnan
Senior Manager
Deloitte Consulting LLP
vedramakrishnan@deloitte.com

Vaibhav Jain
Manager
Deloitte Consulting LLP
vaibhjain@deloitte.com

Siddhant Nagar
Sr Consultant
Deloitte Consulting LLP
sidnagar@deloitte.com

Gina Fratarcangeli
North America GSI Leader
Google Cloud
ginafrat@google.com

David Zhu
Tech Architect Leader
Deloitte Consulting LLP
davzhu@deloitte.com

Pavan Kamarthi
Tech Architect
Deloitte Consulting LLP
pkamarthi@deloitte.com

ACKNOWLEDGEMENTS

LEARN MORE

Visit our Website
Learn more about Deloitte's agentic AI solutions:
https://www.deloitte.com/googlecloud/ai-and-genai

END NOTES

1 Deloitte, State of AI in the Enterprise: The untapped edge, January 2026, p. 5,
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

2 Deloitte, State of AI in the Enterprise: The untapped edge, January 2026, p. 4,
https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

3 Stanford University,, The Enterprise AI Playbook: Lessons from 51 Successful Deployments, April 2026, p.11,
https://digitaleconomy.stanford.edu/app/uploads/2026/03/EnterpriseAIPlaybook_PereiraGraylinBrynjolfsson.pdf

About Deloitte