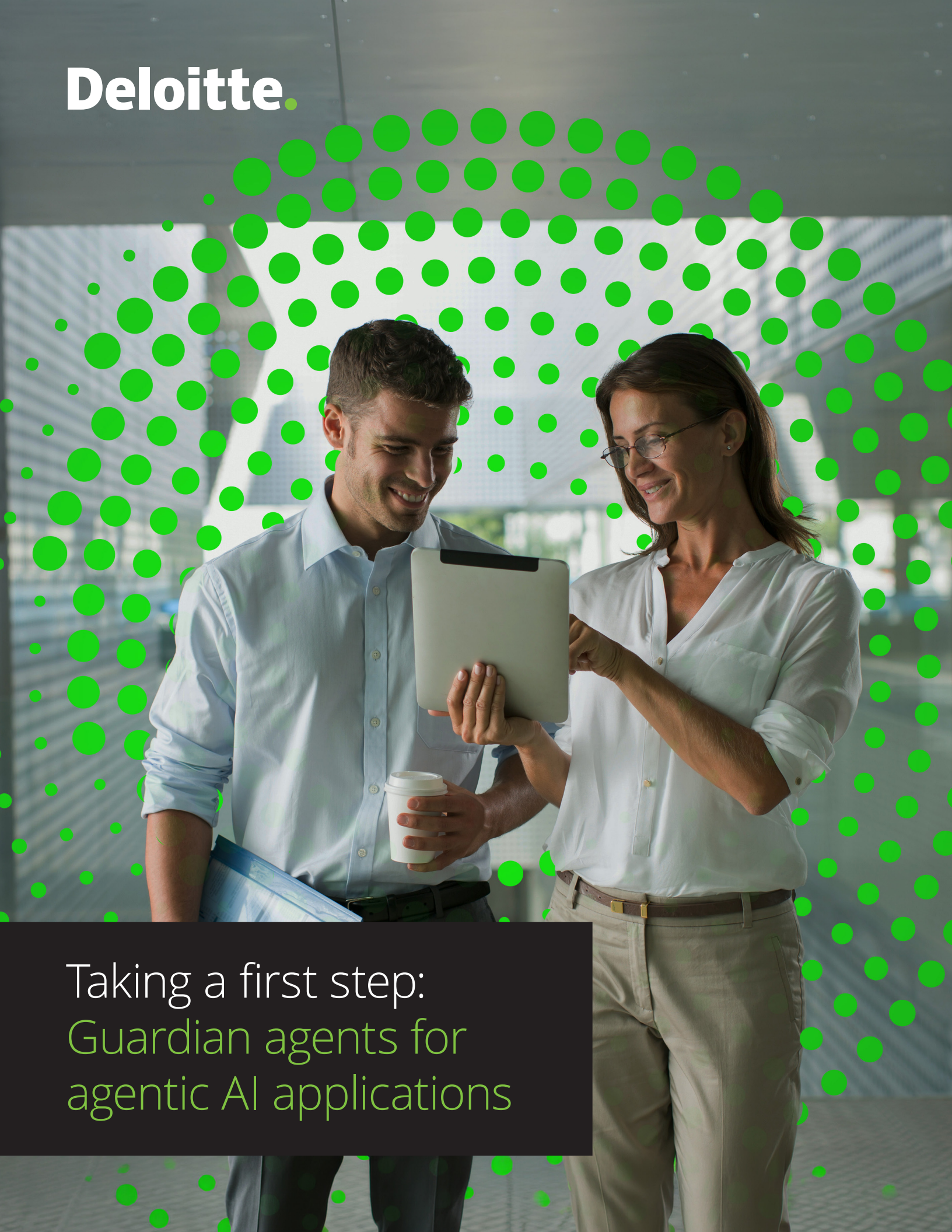
Deloitte.



Taking a first step: Guardian agents for agentic AI applications

Taking a first step: Guardian agents for agentic AI applications

Correct answers impress us with their precision, seemingly correct answers still sway us with their fluency, and incorrect answers often win our leniency—since “always right” was never the standard.

This fascination pattern—marked by the willingness to be impressed and to forgive—showcases both the admiration held for AI's potential and the acceptance of its imperfections. However, such forgiveness can have serious repercussions: As Generative AI models and agentic AI systems rapidly evolve to drive efficiencies across business operations, they can introduce inherent risks that organizations must proactively anticipate and manage thoughtfully. The risks traditionally associated with Generative AI are amplified in agentic AI systems due to autonomous agent chains that expand attack surfaces (i.e., an increase in the number of points where the system can be targeted and compromised) and magnify the impact of failures. Additionally, agentic AI also introduces unique risks like cascading hallucinations, lateral propagation of malicious behaviors, and unauthorized real-world actions by rogue agents. As the AI ecosystems grow in complexity and scale, managing such complex autonomous workflows requires dynamic monitoring, real-time anomaly detection, and remediation mechanisms to maintain safety and trust.

To address these challenges, the concept of guardian agents has recently emerged as a potential approach in AI oversight. Guardian agents are specialized AI systems designed to oversee, monitor, and manage other AI agents.

As Generative AI models and agentic AI systems ***rapidly evolve to drive efficiencies*** across business operations, they can introduce inherent risks that organizations must proactively anticipate.

Depending on the complexity or ambiguity of an agentic AI use case, guardian agents can be designed either to escalate to human oversight or to act autonomously. These agents can be embedded directly into the agentic AI systems during their development as integral components of their functionality, or they can operate independently as stand-alone entities added post-implementation to monitor and oversee the agentic AI systems—either way complementing native governance capabilities with adaptive, real-time intervention.

With the expanding deployments of agentic AI systems—characterized by autonomous, multistep, and context-aware workflows—the traditional, often-used rule-based guardrails focused on predefined risk categories and boundaries also require an evolution to meet new demands from agentic AI systems. Guardian agents potentially represent this next step by introducing an intelligent layer of supervision that continuously validates AI actions, actively guards against threats, and reduces the need for manual oversight. These agents integrate multiple evaluators capable of real-time monitoring, human-in-the-loop escalation, and autonomous correction of issues as they arise.

Gartner and industry analysts project that guardian agents will account for 10% to 15% of the agentic AI market by 2030.[2] This growth reflects the rising demand among enterprises for automated trust, robust risk management, and secure agent oversight solutions as both the number and complexity of AI agents continue to soar. Additionally, multiagent systems are expected to be used in 70% of AI applications by 2028,[3] making guardian agents important for anomaly detection in generated outputs and for effectively managing inter-agent interactions. The next three to five years will likely be defined by the race to embed guardian agents into core enterprise multiagent systems. Success in this market will hinge on achieving secure, transparent automation that blends powerful agentic AI monitoring with human oversight, all while handling the complexity of global compliance and integration at scale.

Deloitte is successfully implementing a guardian agent framework focused on real-time review of results for a pilot agentic AI solution.

"They (guardian agents) function as both AI assistants (supporting users with tasks like content review, monitoring and analysis) and as evolving ***semi-autonomous or fully autonomous agents*** (capable of formulating and executing action plans as well as redirecting or blocking actions to align with predefined agent goals)."[1]

Guardian agents *implementation*

- ***Use case:*** LoanOps is a multiagent system that automates the customer loan inquiry process end to end, from parsing customer emails to verifying attached documents, retrieving financial data, performing payoff retrievals, and drafting client responses; thus, enabling rapid, autonomous handling of high-volume customer loan inquiries.
- ***Risk:*** This autonomous multiagent operation in the LoanOps tool, similar to any agentic AI solution, could have risks such as hallucinated outputs, misinterpretation of intent, reasoning errors, and lack of intermediate validation, making it necessary to deploy guardian agents to ensure continuous verification and trustworthiness.
- ***Guardian agent design and implementation:*** For the LoanOps tool, we utilize Gartner's classification of guardian agents with three primary types: Monitors, Reviewers, and Protectors (figure 1). Specifically, Monitors oversee workflows, Reviewers validate outputs, and Protectors respond promptly to rogue agent activities. These guardian agents are integrated within the LoanOps solution to deliver ongoing, real-time monitoring and proactive intervention at every stage—detecting anomalies, ensuring compliance, and maintaining process integrity. This approach results in faster, more reliable responses while reducing the operational burden on the team.

Monitors

Observe and track AI and agentic actions for ***human or AI-based follow-up.***

Reviewers

Identify and review AI-generated output and content for ***accuracy and acceptable use.***

Protectors

Adjust or block AI and agentic actions and permissions using ***automated actions during operations.***

The guardian agents are designed around three foundational principles that guide their behavior, decision-making, and scope of oversight (figure 2).

Multiple guardian agents were implemented to oversee key workflow stages, including Inquiry Intent Guardian (detects malicious intent), Inquiry Summary Guardian (validates summaries), Payoff Guardian (checks payoff data), Document Analyzer Guardian (verifies documents), Compliance Validation Guardian (confirms output integrity), and Response Validation Guardian (sanitizes final responses).

Testing on a sample data set of emails to customers (LoanOps output) revealed that the LoanOps tool's rate of correct and complete responses improved from 43.8% without guardian agents to 81.25% with their inclusion, while incorrect outputs dropped from 25% to zero. Although there was a slight increase in latency, it remained within acceptable limits, achieving a practical balance between accuracy and efficiency.

Challenges and future considerations: While guardian agents autonomously monitor, audit, and intervene in real time to help reduce compliance, operational, and security risks, as well as threats, deploying them within agentic AI solutions brings technical and operational challenges—such as development skill shortage, integration complexity, balancing flexibility and oversight, ongoing supervision of guardian agents themselves, and managing expectations amid vendor hype for autonomous AI governance. Addressing these issues is critical for organizations to maximize their efficacy and value in practical use.

Figure 2: Guardian agents: Design principles

Scope of oversight

Defines which LoanOps workflow output that a guardian agent is supervising (e.g., payoff data, email summaries, or generated responses).

Evaluation focus

Specifies the evaluation criteria or rule sets used to assess those outputs, such as hallucination, completeness, and compliance.

Response mechanism

Determines the corrective or escalation response based on the evaluation result's severity and business impact.

Key takeaways

1. What are guardian agents? Guardian agents are specialized semi-autonomous or fully autonomous agents designed to supervise, monitor, and manage other AI agents to proactively address inherent risks in rapidly evolving agentic AI systems, such as cascading failures, rogue behavior, and security threats.

2. Why are they important in AI oversight? As autonomous agentic workflows are becoming more complex and significant in business operations, the often-used rule-based guardrails require an evolution to meet new demands, leading to emergence of guardian agents, which introduces an intelligent layer for dynamic supervision.

3. Deloitte's guardian agent implementation in its LoanOps solution: Deloitte has utilized Gartner's classification of guardian agents—Monitors, Reviewers, and Protectors—throughout the LoanOps workflow to deliver real-time monitoring, validation, and intervention at every stage. This includes specific agents for intent detection, summary validation, payoff data checking, document analysis, compliance, and response validation—improving both accuracy and process integrity.

4. Impact of guardian agents in the LoanOps tool: Testing revealed LoanOps' rate of correct and complete responses improved from 43.8% without guardian agents to 81.25% with their inclusion, while incorrect outputs dropped from 25% to zero. Although there was a minor rise in latency, it stayed within acceptable limits, effectively balancing accuracy and efficiency.

5. Challenges in deploying guardian agents: Some key challenges include difficulties in integration, scarcity of skilled development resources, balancing flexibility with oversight, the need for continuous supervision of guardian agents themselves, and managing expectations amid industry hype about autonomous AI governance. Successfully addressing these challenges is critical to realize the full value and efficacy of guardian agents.

Endnotes

1. Avivah Litan and Daryl Plummer, *Guardians of the future: How CIOs can leverage guardian agents for trustworthy and secure AI*, Gartner, April 24, 2025.
2. Gartner, "Gartner predicts that guardian agents will capture 10–15% of the agentic AI market by 2030," press release, June 11, 2025.
3. Litan et al., *Guardians of the future: How CIOs can leverage guardian agents for trustworthy and secure AI*.

Meet the *team*

Scott Holcomb
Deloitte US Trustworthy AI Leader
Principal
Deloitte Consulting LLP
sholcomb@deloitte.com

Clifford Goss
Deloitte US Trustworthy AI Leader
Partner
Deloitte & Touche LLP
cgoss@deloitte.com

Prakul Sharma
AI and Data Leader
Principal
Deloitte Consulting LLP
praksharma@deloitte.com

Amandeep Singh
Trustworthy AI and Risk Leader
Specialist Leader
Deloitte & Touche LLP
amandeesingh@deloitte.com

Tanmay Shrivastava
Trustworthy AI
Go-To-Market Leader
Senior Manager
Deloitte Consulting LLP
tshrivastava@deloitte.com

Sahil Bansal
AI and Data, Manager
Manager
Deloitte
shahibansal@deloitte.com

Vishnu Santhana
Model Risk, Manager
Manager
Deloitte
viskrishnan@deloitte.com

Deloitte.