Whitepaper

Scaling AI agents without rework: a guide for product and engineering leaders

Learn how to scale AI agents across systems without creating new authentication and access overhead for every workflow, integration, or release.

Contents

2 The market shift toward AI agents

2 The identity gap slowing AI scale

3 Understanding AI agents and the systems behind them

5 The four identity challenges behind scalable AI agents

7 The identity framework behind scalable AI agents

10 Building identity systems for the agentic era

The market shift toward AI agents

The global generative AI market is projected to grow from USD 67.18 billion in 2024 to USD 967.65 billion by 2032, exhibiting a compound annual growth rate (CAGR) of 39.6% during the forecast period.

Source: Fortune Business Insights, 2025

The AI agents market is expected to expand from USD 5.1 billion in 2024 to USD 47.1 billion by 2030, reflecting a CAGR of 44.8%.

Source: Fortune Business Insights, 2025

A survey by IBM reveals that 59% of enterprises already working with AI intend to accelerate and increase investment in the technology, reflecting a strong commitment to expanding AI capabilities.

Source: IBM, 2024

The identity gap slowing AI scale

AI agents are moving from experiments to production systems faster than most organizations can adapt. Gartner predicts that 40% of enterprise applications will embed task-specific AI agents by the end of 2026, an 8x jump from 2025. At the same time, 88% of AI agent projects never make it out of the pilot phase, with integration complexity and identity governance cited as major blockers.

As AI agents begin operating across APIs, data systems, and business workflows, teams often end up creating new authentication, authorization, and access patterns for every use case. Security reviews and identity implementation can add months to deployment timelines, forcing engineering teams to spend more time managing integrations, permissions, and token flows than building differentiated product experiences.

Without shared identity infrastructure, organizations struggle to scale AI agents consistently across systems, leading to fragmented workflows, rising maintenance overhead, and slower delivery over time.

88%

of AI agent projects never make it out of the pilot phase

Understanding AI agents and the systems behind them

What Are AI Agents?

AI agents are autonomous software systems that use large language models (LLMs), machine learning (ML), and APIs to perform tasks without direct human intervention. Unlike traditional software, AI agents can interpret and respond to natural language, analyze real-time data, and take action across systems on behalf of users.

These capabilities make AI agents powerful tools for automation, productivity, and customer engagement. They also introduce a new operational challenge: most identity systems were designed around human users interacting through browsers and manual workflows, not autonomous software operating continuously across APIs and services.

As organizations scale AI agents across applications, data systems, and workflows, identity becomes a core infrastructure layer that determines whether teams can move quickly without creating repeated engineering work.

Why Enterprises Are Adopting AI Agents

Organizations across industries are deploying AI agents to improve efficiency, reduce operational overhead, and accelerate customer and employee workflows. But AI agents are also changing how people interact with technology in everyday life. Instead of navigating apps, forms, and disconnected systems manually, users increasingly expect AI systems to handle tasks for them in the background.

AI agents can already help people book travel, purchase concert tickets, manage schedules, coordinate customer support issues, and complete multi-step workflows across applications without requiring constant user input. Inside the enterprise, they are becoming a new execution layer that helps teams move faster and automate repetitive work.

Use cases continue to expand as organizations grow more comfortable integrating AI into day-to-day operations. Some of the most common applications today include:

- **Customer support automation:** AI agents act as frontline support, answering common questions, resolving issues, and escalating complex cases to human agents. This reduces costs and response times while improving the customer experience.
- **Sales and marketing acceleration:** AI agents can qualify leads, draft personalized outreach, and assist with campaign execution using real-time CRM data. They help teams move faster with the headcount they already have.
- **Personal task coordination:** AI agents can help users manage travel plans, coordinate calendars, compare products, monitor ticket availability, and complete tasks that normally require jumping between multiple applications and services.
- **Internal productivity and IT support:** Within the enterprise, AI agents help employees reset passwords, request access, answer HR or policy questions, and even troubleshoot code, offloading routine tasks from IT and operations teams.

As AI agents become more autonomous, organizations need a consistent way to manage identity, access, and runtime permissions across systems. Without a shared identity infrastructure, every new agent or workflow introduces additional integration and maintenance work.

Different stakeholders across the organization experience this challenge differently:

- **CTOs (Chief Technology Officers)** need AI systems that scale across existing infrastructure without creating long-term maintenance and scaling challenges.
- **CPOs (Chief Product Officers)** need teams to ship AI-powered experiences quickly without introducing repeated integration work or approval bottlenecks.
- **CIOs (Chief Information Officers)** need governance and visibility across AI-driven workflows without slowing delivery or creating disconnected identity silos.

The four identity challenges behind scalable AI agents

To scale AI agents efficiently, organizations need a consistent way to handle authentication, API access, long-running workflows, and runtime authorization across systems. Without that foundation, engineering teams can quickly accumulate identity rework that slows deployments and increases maintenance overhead.

As AI adoption grows, organizations repeatedly encounter four core identity challenges that determine whether AI systems scale cleanly across applications, APIs, and connected workflows.

What is RAG?

RAG (Retrieval-Augmented Generation) optimizes AI accuracy by grounding responses in your specific data. It retrieves relevant context from your documents in real-time, helping ensure outputs are factually sound and up-to-date.

Why is token vaulting important?

Centralized token management helps secure the entire AI lifecycle. By automating token exchange and refreshes, you eliminate the need for custom infrastructure while helping ensure seamless, more secure API integrations.

01. User authentication and delegated authority

AI agents need a reliable way to understand who the user is, what actions they are allowed to take, and what context should be carried across systems and workflows. Without a consistent identity model, organizations create attribution gaps where automated actions appear disconnected from the user who initiated them.

As AI agents become more autonomous, teams need delegated authority that links agent actions back to verified user identities. This allows agents to operate on behalf of users while maintaining clear accountability, auditability, and consistent access across applications and services.

Without this foundation, organizations often end up reconciling logs, permissions, and approval histories manually as AI workflows expand.

02. Calling APIs across connected systems

As organizations expand AI workflows across tools, APIs, and MCP servers, managing credentials and integrations consistently becomes increasingly difficult. Without centralized token management, teams often rely on static credentials, duplicated integrations, or custom token handling that creates operational friction as adoption grows.

Token vaulting provides a more scalable model by handling token storage, refreshes, and exchanges centrally. Instead of rebuilding authentication flows for every new integration, teams can standardize how agents connect across systems in a more secure manner while reducing the overhead of maintaining credentials manually.

As the number of connected tools grows, this approach allows organizations to scale integrations without creating additional identity overhead for every new workflow.

03. Long-running workflows and asynchronous authorization

AI agents increasingly operate across workflows that take minutes, hours, or even days to complete. Traditional approval models were designed around users actively interacting with applications, not autonomous systems operating continuously in the background.

Without asynchronous authorization, organizations are forced into a tradeoff: constantly interrupt users for approvals or grant agents overly broad permissions just to keep workflows moving.

Asynchronous authorization allows organizations to automate routine workflows while reserving human approvals for high-impact actions. Agents can pause when additional authorization is needed, request approval through a separate channel, and continue execution once approved.

04. Fine-grained authorization for RAG and runtime access

AI agents often retrieve information from multiple systems in order to generate responses, complete workflows, or take action on behalf of users. As organizations scale RAG architectures, broad access permissions become increasingly difficult to manage safely and consistently.

Traditional role-based access models were not designed for AI systems operating dynamically across large volumes of enterprise data. Without fine-grained authorization, organizations often fall back on hard-coded permissions and manual filtering logic that becomes difficult to maintain as systems grow.

Fine-grained authorization allows organizations to control exactly what data, actions, and workflows an AI agent can access at runtime. Permissions can be scoped to the specific task, user, or resource involved rather than granting broad standing access.

As AI systems become more interconnected, these identity challenges compound across applications, APIs, workflows, and enterprise data systems. Without consistent identity patterns, teams often fall back on fragmented integrations, custom token handling, and one-off authorization logic that becomes increasingly difficult to maintain over time.

The identity framework behind scalable AI agents

That's why we built Auth for AI Agents. It provides a reusable, standardized identity architecture designed to help organizations scale AI systems across applications, workflows, APIs, MCPs and connected services.

The following blueprint focuses on four common identity bottlenecks that emerge as AI systems become more autonomous and interconnected.

01. Delegated authority

The problem: Attribution gaps create friction at scale

When AI agents operate without a clear link to user identity, organizations lose visibility into who initiated actions across workflows and systems. Teams often end up manually reconciling logs, permissions, and approvals just to trace activity across AI-driven processes.

The solution: Authentication + delegated authority

Auth for AI Agents links AI actions back to verified user identities so agents can safely operate on behalf of users across applications and workflows while maintaining accountability and consistent access propagation.

The result: Scale AI workflows without audit overhead

By connecting AI activity directly to user identity, organizations reduce manual reconciliation work and maintain clearer operational visibility as AI adoption grows.

02. Token vault

The problem: Credential sprawl slows everything down

As AI agents connect across more tools and APIs, teams often end up managing fragmented credentials, duplicated integrations, and custom token flows that become increasingly difficult to maintain.

The solution: Managed credential infrastructure

Auth for AI Agents centralizes token storage, refreshes, and exchanges so teams can standardize API access across applications, MCP servers, and connected systems without rebuilding integration logic repeatedly.

The result: Scale integrations without added friction

As organizations expand AI workflows across dozens of tools and services, the operational effort required to support new integrations drops significantly.

03. Async authorization

The problem: Approvals don't scale with AI

Traditional approval flows force users to stay actively involved in long-running workflows. Organizations either interrupt teams constantly for approvals or grant overly broad permissions just to keep workflows moving.

The solution: Automated approvals with human oversight

Auth for AI Agents supports asynchronous authorization flows where AI agents can pause, request approval when necessary, and continue execution once authorization is granted.

The result: Scale workflows without approval bottlenecks

Teams can automate routine operations while reserving human oversight for high-impact actions, reducing manual interruptions without sacrificing operational control.

04. Fine-Grained Authorization

The problem: Hard-coded permissions don't scale

AI agents operating across enterprise systems often rely on static roles and broad permissions that become increasingly difficult to manage consistently as workflows expand.

The solution: Fine-grained authorization

Auth for AI Agents applies runtime authorization controls that scope access to the specific task, resource, and user context involved in an interaction.

The result: Scale access without engineering drag

Teams can support RAG-based systems and connected AI workflows without building custom authorization logic for every application and dataset.

05. MCP server authorization

The problem: MCP connections create new identity gaps

As AI agents retrieve context from MCP servers across applications and services, organizations need a consistent way to manage identity propagation, permissions, and runtime access across connected environments. Without standardized authorization, teams often fall back on fragmented access models and custom permission handling for every MCP integration.

The solution: Authorization for MCP servers

Auth for MCP helps organizations standardize authentication and authorization across MCP-based interactions, allowing AI agents to retrieve context while maintaining consistent identity and access controls.

The result: Teams can expand AI-driven workflows across connected tools and data systems without rebuilding identity logic or creating disconnected access patterns for every MCP server.

Building identity systems for the agentic era

As AI agents become more integrated into business workflows, organizations need identity systems that can scale without creating repeated engineering work. Teams should focus on:

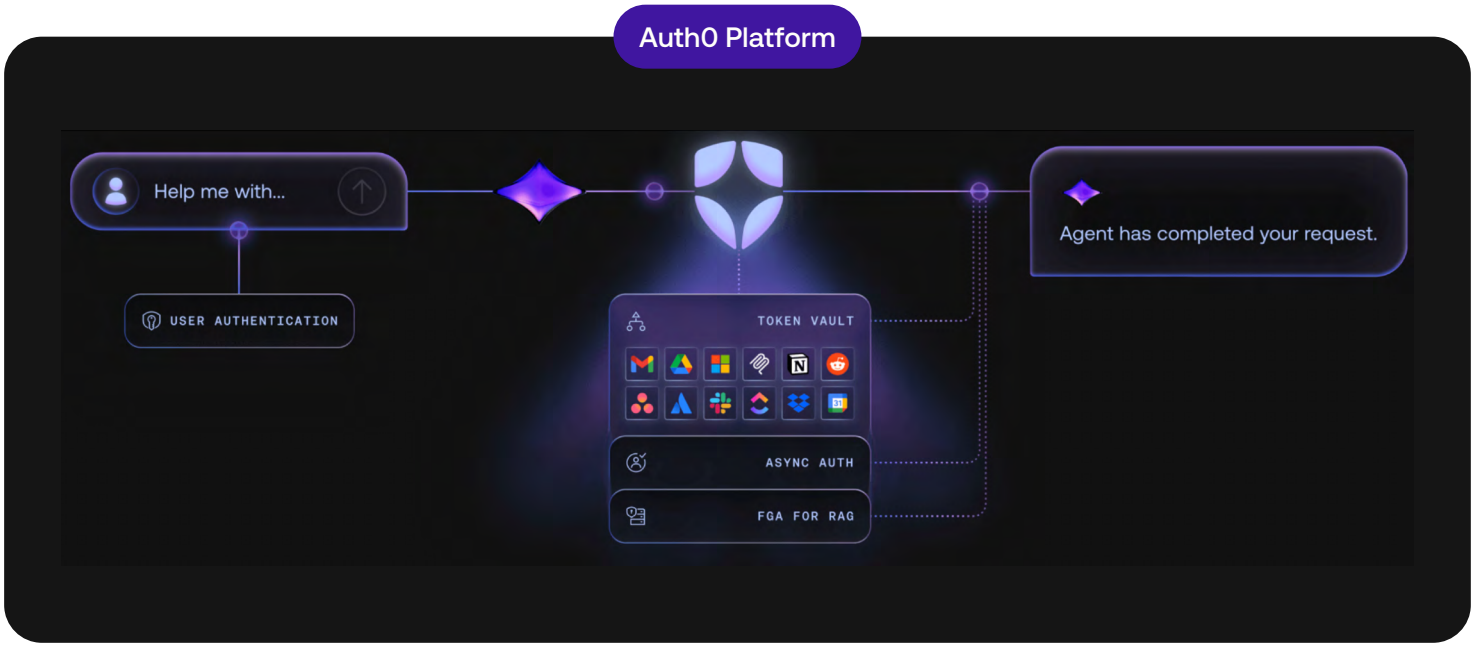
- **Standardizing identity across AI systems:** Reduce fragmentation by using consistent authentication and authorization patterns across applications, APIs, and AI agents.
- **Replacing custom integrations with reusable identity infrastructure:** Avoid rebuilding token handling, approval logic, and access controls for every new workflow or integration.
- **Supporting long-running and autonomous workflows:** Enable AI agents to operate across asynchronous processes while maintaining clear user context and approvals when needed.
- **Applying fine-grained access controls across systems:** Help ensure AI agents only access the data, actions, and workflows required for a specific task, even across connected environments.
- **Securing MCP servers and context flows:** Standardize authentication and authorization across MCP servers so AI agents can retrieve context from connected applications and data systems without creating fragmented access patterns or inconsistent permissions.
- **Building identity systems that can scale with AI adoption:** As AI use cases expand, organizations need identity infrastructure that supports new agents and workflows without increasing operational overhead.

AI is changing how applications, systems, and users interact. Over the next decade, organizations will rely on growing ecosystems of AI agents operating across workflows, APIs, MCP servers, and enterprise data systems.

As these environments become more connected and autonomous, identity becomes the coordination layer that allows teams to scale AI safely and consistently without rebuilding access controls, integrations, and approval logic for every new use case.

Organizations that treat identity as reusable infrastructure will be able to move faster, scale AI systems more efficiently, and support new business workflows without creating fragmented systems that become harder to maintain over time.

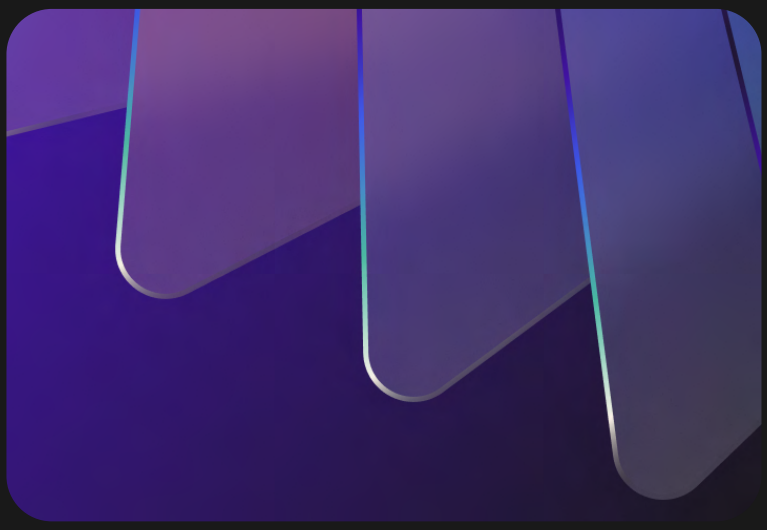
How Auth0 for AI Agents Works



With Auth0, teams can move identity out of custom application code and into a reusable infrastructure layer built for AI-driven systems. Instead of stitching together authentication, authorization, and approval logic for every workflow, organizations can scale AI agents faster with a consistent identity foundation across applications, APIs, and services.

Learn more about building scalable AI systems with Auth0 at auth0.com/ai.

About Auth0
Auth0® provides a modern identity platform that helps organizations secure and scale applications, APIs, and AI-driven experiences. Flexible enough for complex enterprise environments and simple enough for fast-moving development teams, Auth0 enables organizations to manage identity across users, systems, and non-human actors without rebuilding identity infrastructure for every new use case. Auth0 is a part of Okta, Inc., The World's Identity Company™.



Scaling AI agents without rework: a guide for product and engineering leaders

okta
The World's Identity Company™

Okta Inc.
100 First Street
San Francisco, CA 94105
info@okta.com
1-888-722-7871