The agentic AI dilemma

Balancing speed of adoption with security and scale

Enterprises adopting artificial intelligence (AI) have to walk a fine and difficult line between excessive haste and healthy ambition.

AI technology is developing quickly, and enterprises may fear falling behind. But AI adoption is also risky and expensive. Already, 41% of organizations admit to adopting generative AI too quickly, according to PagerDuty research.

Agentic AI stands to up these stakes. AI agents are uniquely susceptible to cost overruns, user experience problems, and breaches due to their architectural complexity and heavy use of computing power.

But none of these risks are likely to slow enterprises down. Since agents can autonomously execute complex sequences of tasks, many technology leaders predict big gains in productivity and agility. According to that same PagerDuty survey, 95% of executives plan to adopt agents even faster than they adopted generative AI.

Therefore, enterprises need the best of both worlds — the ability to capture AI agentic value quickly while also minimizing the risks speed often brings. They can strike this balance with:

Network
Thinking globally about agentic AI architecture

Pricing
Finding flexibility in the compute pricing model

Security
Prioritizing automation and service integration

Consideration #1: Thinking globally about agentic AI architecture

Even the most innovative AI application needs to deliver a fast, reliable user experience.

Key takeaways

- 41% of companies admit they adopted generative AI too hastily.
- 95% of companies plan to adopt agentic AI even faster than GenAI.
- Agentic AI adoption demands both speed and control.

Agentic applications may struggle in this regard. They are likely to connect a variety of external applications and databases, which allow them to pull data and perform tasks. If those applications and databases live on distant infrastructure (e.g., an on-premises data center), latency, and other delays can compound quickly, especially during usage spikes from distributed user bases.

To help agentic applications scale without adding additional effort, enterprises should ask themselves: how much of the application (sub-agents, large language models (LLMs), Model Context Protocol (MCP) servers, etc.) can be made to run globally? Not just in a massive data center in a specific region, but at the network edge. The shorter the distance (in terms of both geographic distance and network hops) between the application, its users, and the external resource it connects to, the more confidence an enterprise can have before a rapid launch.

Consideration #2: Find flexibility in your compute pricing model

From a cost perspective, the ability to scale AI agents down is just as important as the ability to scale up. AI computation is uniquely expensive compared to other types of applications. Enterprises need to be cautious of investing in compute resources that ultimately go unused.

Unfortunately, many hyperscalers' pricing models complicate this issue. On a micro level, many charge for "wall clock time" — much of which an agent may spend waiting on a response from component LLMs, MCP servers, or APIs.

Meanwhile, on a macro level, hyperscalers often charge for large blocks of computing capacity, whether or not all of that capacity is actually used. And with agentic applications, the answer is often "not," because AI inference draws on compute resources unpredictably.

Some enterprises may have plenty of unused compute capacity to burn. But if they do not, or if they are unsure how much usage their agents will get, they should explore alternatives that offer more flexible pricing at a macro and micro level, and which do not force enterprises into wild guesses.

Consideration #3: Prioritizing security integration and automation

Developing any sort of application too quickly is risky, and agentic AI is no different. The risk manifests in:

- **Interconnections:** As mentioned, a single agentic application may interconnect with many external applications and databases via APIs or MCP servers. Organizations may struggle to ensure agents do not access, read, or write sensitive data beyond their intended scope. For example, an estimated 30% of enterprise APIs are already not being tracked.
- **Excessive agency:** In addition to authentication and authorization, organizations need to constrain the volume of actions an AI agent can execute within defined timeframes. Otherwise, agents may engage in hazardous repetitive behaviors, like dispatching identical email responses to the same recipient thousands of times.
- **LLM security risks:** A single agentic application may use multiple LLMs, which still have uncertain data security practices. A DeepSeek breach was revealed less than two weeks after its wildly popular 2025 launch. By 2027, Gartner predicts, 40% of all breaches will stem from cross-border generative AI misuse alone.
- **Standard web application attacks, e.g., distributed denial-of-service (DDoS):** Time- and resource-strapped developers may not always put the proper protections in front of the customer-facing agents they launch.

Various services already exist to protect these different aspects of agentic AI functionality. But applying a bunch of new, isolated tools is no way to accelerate development. Many organizations' IT environments are already extremely complex.

Instead, enterprises should prioritize integration and efficiency for their security services. This means — for example — an API gateway, AI gateway, AI-focused firewall, and Zero Trust security services that are easily interoperable, manageable from a single interface, and draw on a single, strong body of AI-inclusive threat intelligence.

Automation is another important consideration. Ideally, critical tasks, like discovering agentic APIs and applying web application protections, should not rely solely on human effort.

Cloudflare's connectivity cloud helps enterprises adopt agents both quickly and safely

Cloudflare's connectivity cloud is a unified platform of connectivity, security, and developer services powered by a global cloud network.

Many of these services support fast and safe agentic AI adoption. For example, Cloudflare's network spans over 330 global cities, and the majority of our AI services can operate in every network location. This lets organizations run agents and LLMs close to end users.

In addition, Cloudflare does not make organizations pay for unused graphics processing units (GPUs), or for the time agents spend waiting on APIs or external LLMs.

And on the security side, Cloudflare unites on a single platform the security services needed to secure customer-facing agents, developer access to those agents, and connections between agents. This includes a Firewall for AI, Zero Trust services, and integrations with agentic authentication services. What is more, Cloudflare protects 20% of all web properties —including more than 80% of the 50 largest AI companies — and automatically applies the resulting threat intelligence to all of our security services.

Learn more about Cloudflare's AI services in our ebook, "The enterprise guide to securing and scaling AI."