

What AI Agents Actually Deliver

How to Find Reliable Value Before Expanding Autonomy

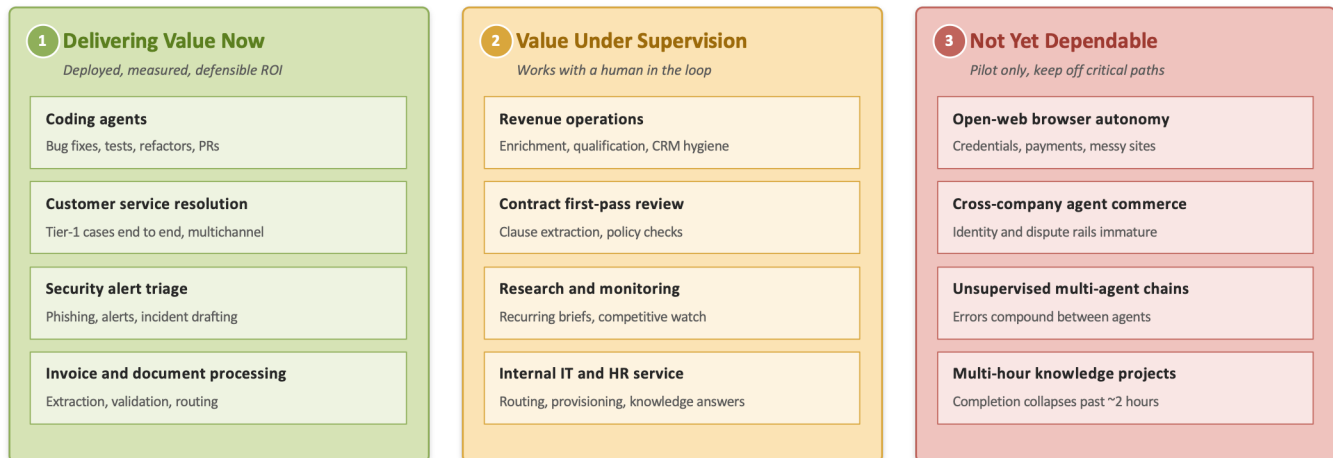
Our paper [From Chatbots to Digital Operators](#) described AI agents as systems that can plan, use tools, and carry multistep work toward an outcome with limited supervision. This paper focuses on value. Agents create value when work is observable and bounded, with a human path for exceptions. Model capability cannot compensate for poor workflow design.

Two findings show why. In a survey of more than 500 technical leaders, 80% reported measurable returns from deployed agents. Yet Gartner expects more than 40% of agentic projects to be canceled by the end of 2027, while deployment remains limited to 17% of organizations despite more than 60% planning to deploy within two years. Together, these findings show that early returns and long-term viability are different measures. Long-term value depends on choosing workflows an organization can govern and defining outcomes that can be measured.

Value concentrates in work you can verify

Valuable work repeats inside systems that record each action, and a person can quickly accept the output or handle an exception. Outside those conditions, an agent is still a research project with an operating budget.

Where AI Agents deliver Value



As of August 2026, placement reflects public evidence and observed production behavior, not vendor claims.

Coding proves the value and exposes the cost

Software engineering shows the strongest evidence of production use. OWASP tracked 53 agentic projects and found 28 were coding agents. The five fastest-growing tools were all in that group. Cursor reached \$2 billion in annual recurring revenue in February. At Uber, the share of engineers classified as agentic coding users rose from 32% in February to 84% by March 2026.

The economics require equal attention. Uber exhausted its annual AI budget in four months. Typical consumption runs \$150 to \$250 per engineer each month, with heavy users at \$500 to \$2,000. Coding agents can increase throughput, but they turn part of software delivery into a variable consumption cost. Finance needs a usage model before adoption expands.

Customer service is showing scale

Salesforce agreed in June to buy Fin for about \$3.6 billion. Fin reports resolving roughly 76% of incoming support requests without a person, compared with about 62% for the Agentforce help agent. The acquisition would also contribute more than 30,000 customers. It shows how the market values reliable customer resolution.

Production data points the same way. The share of service organizations using AI agents rose from 39% to 66% in a year. The average number of activated agents per organization nearly tripled, build time fell 53%, and 70% of adopters reported measurable value within 60 days. The index covers organizations already running agents every month, so it shows how quickly committed adopters scale rather than broad market adoption. For leaders, the implication is that service agents can scale quickly once workflows and resolution definitions are standardized.

Quiet workflows offer the best risk-adjusted return

The best return per unit of risk is coming from unglamorous work such as alert triage, invoice extraction, first-pass contract review, and exception routing. St. Luke's University Health Network saved close to 200 hours a month on phishing triage. Rossum reported AP automation at Fugro across 21 countries and about 300,000 invoices a year. Legartis reported dormakaba cutting review time on 100-page contracts by 67%. The business case is repeatability, not novelty.

Open-ended autonomy remains a strategic bet

Browser autonomy remains unsuitable for sensitive workflows. OpenAI retired its standalone Atlas browser on August 9, less than ten months after launch, and moved agentic browsing into ChatGPT. Google discontinued Project Mariner in May and folded it into Gemini. Claude for Chrome grew from about 40,000 installs in December 2025 to more than 10 million by June, while Anthropic still describes browser use as inherently risky. A University of Washington study tested seven agentic browsers and found that four allowed attackers to bypass the same-origin policy that prevents one website from reading another's data.

Cross-company commerce is on a similar path. OpenAI retired its Instant Checkout flow in March and redirected its commerce protocol toward product discovery. Google's competing standard went live in AI Mode with major US retailers in May. The infrastructure is advancing, but the first products built on it have retreated. Preserve optionality, but do not build this year's operating plan around it.

Task length is a key constraint

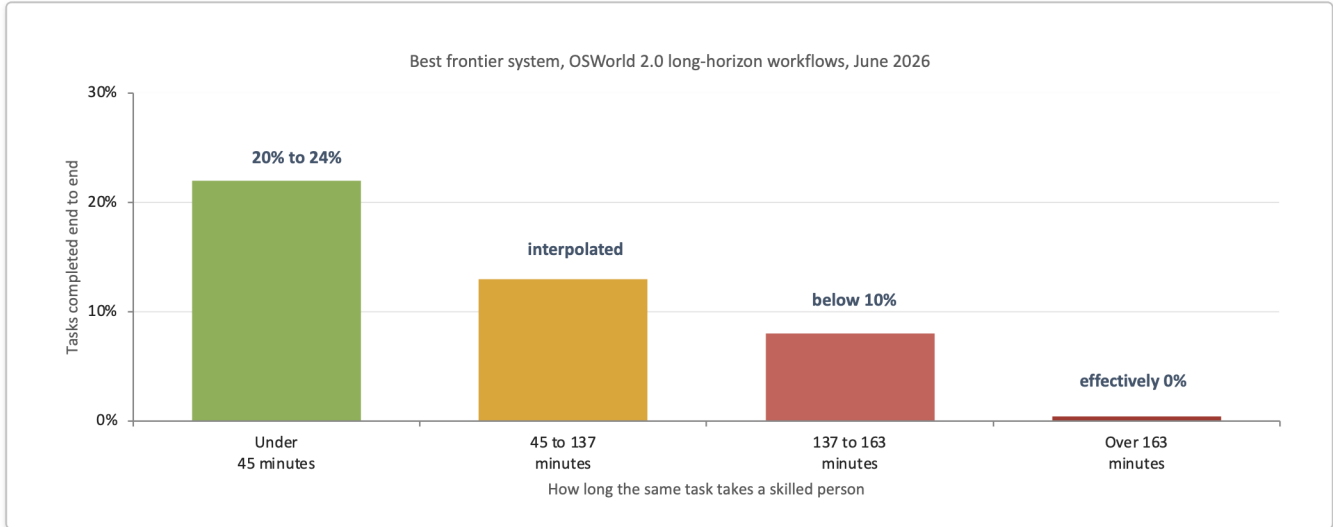
Short computer-use benchmarks now overstate what a business can delegate. The leading verified score on OSWorld is 86.1%, above the human baseline of about 72%. Taken alone, it suggests agents can handle most desktop work.

OSWorld 2.0 tells a different story using longer versions of the same kind of work. Its 108 end-to-end workflows take a skilled person about 1.6 hours to complete and require an average 318 tool calls. The best system completes only 20.6%, with 54.8% partial credit. The difference between apparent mastery and frequent failure is task length.

The failure is not mouse control. Agents lose constraints partway through, overlook earlier information, guess when they should ask, and skip the final check. The core problem is maintaining context and knowing when to stop.

The Reliability Cliff

Agent completion rate falls off sharply with task length.



Reported bins from OSWorld 2.0. The yellow bar is interpolated between two published points

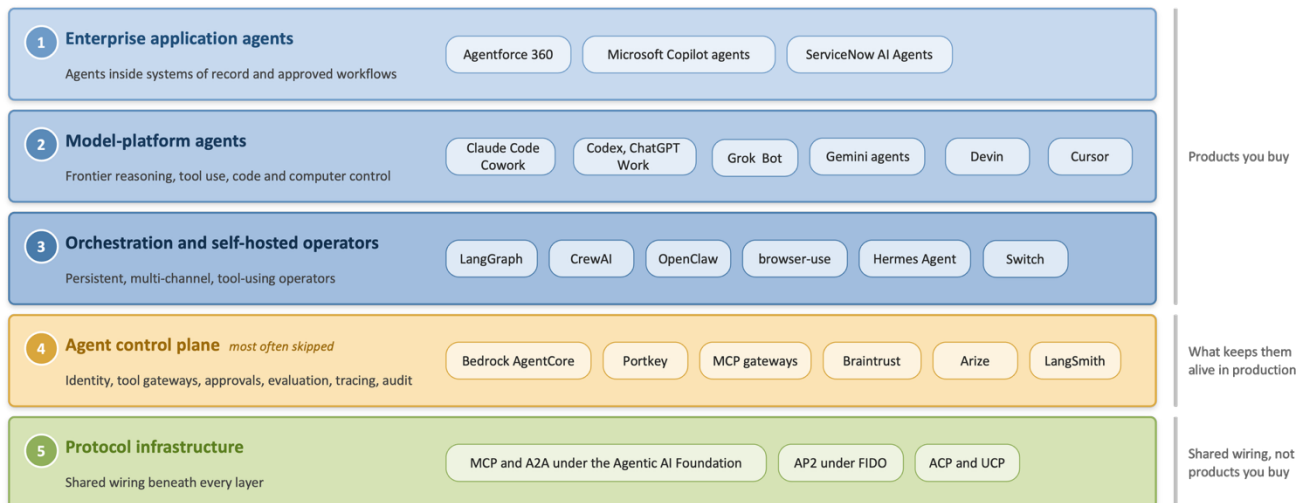
Design around task length. If the longest human version of the work runs past about an hour, do not hand it over whole. Divide the work into segments that each produce evidence a person can inspect, and place approval at those boundaries. Today's business advantage comes from well-designed checkpoints, not from waiting for flawless autonomy.

The Agent Market

Once a workflow passes the reliability test, the next decision is what to buy. Agent vendors occupy four layers above a shared protocol infrastructure. The layer determines which controls are included and which the buyer must build. Many disappointing deployments start by buying at one layer and assuming the protections of another.

Who plays at each layer

Five layers - the lower two are not products you shop for in the same way as the top three.



The Agent control plane governs agent identity, filters tool traffic, routes approvals, evaluates performance, and preserves audit evidence. Forrester began formally evaluating the category this year, and consolidation has started with Cisco's April acquisition of Galileo for Splunk. Installing this layer early turns governance into a shared service. Retrofitting it after dozens of agents turns it into a remediation program.

The protocol infrastructure matured faster than most roadmaps assumed. The Model Context Protocol and Google's Agent2Agent protocol now reside under the same foundation, with more than 250 members that include major cloud and model providers. Payment authorization moved to the FIDO Alliance in April 2026. For buyers, protocol support now belongs in procurement because it lowers switching cost and preserves future choices.

The Agent Market Stack:

Layer	What it does	Representative players	What you get
Enterprise application agents	Agents inside systems of record and approved workflows	Agentforce 360, Microsoft Copilot agents, ServiceNow AI Agents	Inherited identity, permissions, and audit. Easiest to govern.
Model-platform agents	Frontier reasoning, tool use, code and computer control	Claude Code and Cowork, Codex and ChatGPT Work, Grok Bot, Gemini agents, Devin, Cursor	The most capability. More integration and risk work.
Orchestration and self-hosted operators	Persistent, multi-channel, tool-using operators	LangGraph, CrewAI, OpenClaw, browser-use, Hermes Agent, Switch	Maximum flexibility. Safety depends on your configuration.
Agent control plane	Agent identity, tool gateways, approvals, evaluation, tracing, audit	Bedrock AgentCore, Portkey, MCP gateways, Braintrust, Arize, LangSmith	The layer that decides whether the other three survive production.
Protocol infrastructure	Shared wiring beneath every layer	MCP and A2A under the Agentic AI Foundation, AP2 under FIDO, ACP and UCP	Portability. Reduces the cost of changing your mind later.

The advantage now lies beyond the model

The harness now matters as much as the model. Within seven days in April, Anthropic shipped managed agents with sandboxed execution, checkpointing, scoped credentials, and end-to-end tracing. OpenAI updated its agent SDK with a model-native harness, multiple sandbox providers, and a formal split between control and compute. Both competitors converged on the same architecture. Governance and execution infrastructure are now part of the product decision, not an integration task for later.

General work agents have become outcome-first. Anthropic's Cowork reached general availability in April and expanded to mobile and web in July. OpenAI launched ChatGPT Work in July, while Microsoft extended Copilot into many of the same workplace workflows. xAI followed in August with Grok Bot, giving each agent a persistent cloud machine that signs into a company's applications using the user's own credentials. A user describes a result, and the agent plans across connected applications before returning finished files. This could move adoption beyond engineering, but longer autonomous runs make permissions and review design more consequential.

Agents are moving into tools where teams already work. In a single week in August 2026, Slack introduced Slack Code, where teams can direct coding agents and preserve the work as an approval record. GitHub put shared Copilot

agent sessions in Teams into public preview. SandboxAQ released Switch, an open-source, vendor-neutral layer that brings agents built on different frameworks into Slack, Teams, Discord, and other chat platforms. Salesforce had already made Slackbot's MCP client generally available, connecting it with more than twenty partner applications.

Cost per unit of agent work is falling faster than capability is rising. Anthropic released Opus 5 in July at near-frontier performance for about half the price of comparable models. OpenAI shipped its newest family in several sizes, with the smallest matching the largest on many jobs. Labs are also emphasizing agent evaluations and long-context performance over single-turn knowledge scores. Capacity is still finite, so harness traffic is billed at full API rates and plans impose practical limits on agent hours.

Model choice is becoming a routing- rather than an architecture decision. Separate the workflow from the models used to plan, execute, and review it. The best combination will change more than once before most projects reach their second year. That separation protects the workflow investment when prices and model performance move.

The decisions that determine return

Evaluate agents by the cost of an accepted result. Compare that figure with the human cost of completing the same task, including review and correction time. Define "accepted result" in the vendor contract. Vendors may report resolution rates near 71%, while independent estimates place the same performance between 42% and 50%, largely because they count different events as resolutions. That definition can affect your bill more than a model upgrade.

Make identity and disclosure part of launch readiness. With 88% of organizations reporting a confirmed or suspected agent security incident last year, and agents now signing into applications with the user's own credentials, every agent needs its own identity and audit logs it cannot reach. Read the vendor documentation carefully. Customer-facing deployments also need a disclosure policy under EU AI Act requirements that took effect on August 2, 2026.

Decide where agents fit into the way your company operates. Set rules for self-hosted agents before teams begin deploying them independently. OpenClaw, the leading open-source platform reached 368,000 GitHub stars and 12 million downloads despite serious security and configuration problems. Revisit staffing and capacity plans that assume people will perform all engineering and operations work. The earliest gains will come from workflows that are already measured and that leaders can redesign themselves.

Start with work you can measure and verify

The strongest first projects are high-volume workflows with clear policies, accessible data, reversible actions, and an existing human fallback. They create a measurable result without asking the organization to trust open-ended autonomy. Work that depends on tacit judgment or unclear ownership will consume more management attention than it returns.

Before writing prompts, define a short operating contract for each agent. It should set the agent's authority, required evidence, the conditions under which the agent must escalate, and owner. That contract turns acceptable behavior into something teams can test and monitor, while giving finance and auditors a common record. If a workflow cannot produce evidence a person can inspect, it is not ready for production.

Use a limited commercial experiment to test whether customers will pay for completed work instead of access. Scale autonomy only after the workflow has proved its economics and controls.

[West Coast Consulting Group](#) helps organizations turn agentic AI from isolated pilots into operating capability. We help you identify where the economics are credible and design workflows that can scale safely.

Sources

- Anthropic. "The 2026 State of AI Agents Report." 2026. <https://resources.anthropic.com/2026-state-of-ai-agents>
- Yuan, M., et al. "OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks." arXiv:2606.29537, June 2026. <https://arxiv.org/abs/2606.29537>
- BenchLM. "OSWorld-Verified Leaderboard." Aug. 21, 2026. <https://benchlm.ai/benchmarks/osworld-verified>
- Salesforce. "Salesforce Signs Definitive Agreement to Acquire Fin." June 15, 2026. <https://www.salesforce.com/news/press-releases/2026/06/15/salesforce-signs-definitive-agreement-to-acquire-fin/>
- Salesforce. "State of Service: AI Agents Edition" and the 2026 Agentic Enterprise Index. <https://www.salesforce.com/news/stories/ai-service-agents-improve-customer-satisfaction/>
- Forbes. "Uber Burns Its 2026 AI Budget In Four Months On Claude Code." May 2026. <https://www.forbes.com/sites/janakirammsv/2026/05/17/uber-burns-its-2026-ai-budget-in-four-months-on-claude-code/>
- Help Net Security. "Prompt injection still drives most agentic AI security failures in production." June 11, 2026. <https://www.helpnetsecurity.com/2026/06/11/owasp-prompt-injection-ai-security-failures/>
- Linux Foundation. Agentic AI Foundation formation and A2A project announcements. Dec. 2025 and Aug. 2026. <https://www.linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation>
- Gartner. "Over 40% of Agentic AI Projects Will Be Canceled by End of 2027." June 2025. <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
- Aissist. "AI Agent Pricing Benchmark 2026." Aug. 2026. <https://aissist.io/industries/ai-agent-pricing-benchmark-2026>
- No Hacks. "The Agentic Browser Landscape in 2026." July 2026. <https://nohacks.co/blog/agentic-browser-landscape-2026>
- Agentic Commerce Protocol. "What Is the Agentic Commerce Protocol." 2026. <https://agenticcommerceprotocol.info/standards/acp>
- Big Hat Group. "State of OpenClaw 2026: The Enterprise Self-Hosted Agent." May 2026. <https://www.bighatgroup.com/blog/state-of-openclaw-2026-enterprise-self-hosted-ai-agent/>
- Preloop. "The AI Agent Control Plane in 2026." Aug. 2026. <https://preloop.ai/resources/ai-agent-control-plane-2026>
- Augusto Digital. "LLM News August 2026." Aug. 2026. <https://augusto.digital/insights/blogs/monthly-llm-news-august-2026/>