What AI Agents Really Cost, and What Rapidly Growing and Mid-Sized Companies Should Own

How to price agentic AI honestly, avoid lock-in, and build a capability your own people can run

This whitepaper brings together key questions and answers from my executive briefings, conference speaking, participants in my training programs and additional research addressing implementation costs, return on investment, AI model ownership, vendor independence, and the organizational capabilities needed for long-term success.

James Proctor, The Inteq Group | November 2026

KEY TAKEAWAYS

1. An AI agent's total cost includes analysis and design, build and deployment, and ongoing operation. A well-scoped first process can be managed as a defined project, with costs assessed against its business value.
2. AI capabilities can be purchased as services, developed as owned assets, or delivered through hybrid arrangements. The critical question is what rights, controls, and capabilities remain with your organization.
3. Multistep agents can use substantially more computing and tokens than chat tools. Compare pricing models by cost per completed business case, not just cost per token or action.
4. Owned or privately operated AI can be economical at sustained high volume, but the crossover depends on utilization, infrastructure, staffing, and model requirements. Compare full costs over three years.
5. Long-term control requires the skills and rights to operate the capability. Capture expert judgment, define ownership and licensing terms, and make outside support a choice rather than a dependency.

What does an AI agent actually cost?

The cost of an AI agent has three parts, and only one is the technology. For a single, well-scoped first process, it can be managed as a defined project rather than an enterprise-wide program:

- **Analysis and design:** working out how the process runs, which decisions the agent will make, and the rules and limits it works within.
- **Build and deployment:** configuring the agent, integrating it with your systems, and testing it on real cases.
- **Ongoing operation:** infrastructure, monitoring, and periodic updates as models and rules change.

No responsible ROI estimate can be made before examining the process, its transaction volume, and the time and costs it consumes today. Potential gains include shorter cycle times, fewer errors,

lower cost per transaction, and additional capacity without proportional hiring. Actual returns must be measured against a baseline.

Underestimating operating costs is a significant risk. Gartner's 2026 research forecasts that at least half of generative AI projects will exceed their budgets through 2028, driven in part by poor architectural choices and insufficient operational expertise. Gartner also predicts that inference costs per agentic workflow will increase more than fivefold through 2028 as AI systems take on increasingly complex, multistep work. These forecasts reinforce the importance of evaluating total operating costs before moving from pilot to production.

Metered service or owned asset: which deal are you signing?

AI investment decisions generally involve two broad approaches: purchasing AI capabilities as a service or developing capabilities that the organization owns and controls. With a metered service, fees may be based on tokens, requests, actions, or users. An owned capability involves download and configuring an open-weight model and operating it under open-source license. Hybrid arrangements are also possible, combining owned decision logic and applications or managed services.

Neither approach is inherently wrong. Metered services often suit experimentation and occasional or unpredictable work. The key question is what remains when the contract ends. Depending on its terms, a service may leave the organization with outputs and exported data, while a properly structured open-weight owned capability can preserve a working system, documented decision logic, operating knowledge, and rights to maintain and modify it. The same investment can leave an organization with very different levels of control, operational independence, and long-term business capability.

> ***When the contract ends, what do we still have?***

Why do agents change the cost math?

Metered AI services may charge for tokens, requests, actions, conversations, or seats. A token is approximately three-quarters of an English word, although the ratio varies by language and content. Token-based services commonly charge separately for input and output tokens. Individual calls may be inexpensive, but total charges depend on how often and how extensively the capability is used. The important measure is the cost of completing useful business work, not simply the unit price of an AI request.

A chat assistant typically responds to an explicit user request. An agent may work through a queued business case, retrieving records, applying rules, calling tools, and recording results over multiple steps. In its research on multi-agent systems, Anthropic reported that agents in its implementation used approximately four times as many tokens as conventional chat interactions, while multi-agent systems used approximately fifteen times as many. Token consumption varies considerably by application, but the example illustrates how multistep workflows can increase operating costs at scale.

AI vendors are also expanding their pricing options. Salesforce, for example, initially offered Agentforce at $2 per conversation and in 2025 introduced Flex Credits, including pricing equivalent to approximately $0.10 per standard agent action. Its broader offering also includes user-based and

consumption-based arrangements. Whatever the pricing model, convert each quote into an estimated cost per completed business case at current and projected volumes, and account for additional platform, integration, and support charges.

What does it cost to own the capability, and when does it pay off?

Operating an agent built on an open-weight model involves the same broad cost categories: development, infrastructure, and ongoing maintenance. Process analysis is necessary under either delivery model, but development is not always the largest cost component. Some bounded tasks can run on relatively modest infrastructure; others require substantial computing capacity, resilience, security controls, and specialized support. When inference runs on infrastructure the organization controls, costs arise primarily from computing and operations rather than third-party model API charges. The financial advantage depends on workload, utilization, and total operating cost.

The economic crossover between metered AI services and privately operated AI depends on volume, model requirements, infrastructure utilization, and ongoing operating costs. At low or irregular volumes, metered services may be more economical because they avoid dedicated infrastructure commitments. At higher and more consistent volumes, privately operated open-weight open-source models can offer cost advantages, although capacity and operating costs may rise as demand grows.

OECD research published in 2026 illustrates that private hosting may offer little advantage at low utilization but can recover infrastructure costs substantially faster at sufficiently large, steady workloads. Compare both approaches for one representative process over three years, including development, infrastructure, staffing, security, maintenance, expected growth, and potential migration costs.

Stanford's 2025 AI Index documented a more than 280-fold reduction in the cost of querying models achieving approximately GPT-3.5-level performance between November 2022 and October 2024. It also reported a narrowing gap between leading open-weight and closed models on selected benchmarks. These findings broaden the range of work that smaller and open-weight models may handle economically, but they do not establish parity across all tasks or guarantee lower total ownership costs. For bounded operational work, test candidate models against representative cases from your own business, including difficult exceptions, rather than relying only on general-purpose benchmarks.

How do you avoid AI vendor lock-in?

Separate what you want to keep from what you may want to change. Process designs, decision rules, controls, test cases, configurations, and documentation should be captured in terms your organization understands. Design the model and platform to be replaceable where feasible, keep data in formats you control, and avoid unnecessary dependence on proprietary features. Confirm in the implementation agreement which custom deliverables the company owns or can use perpetually, and identify third-party components, source-code access, licensing restrictions, and dependencies that might limit future maintenance or migration.

Then ask each provider the same question: if we changed vendors next year, what would we need to rebuild or replace? The shorter and more predictable the answer, the less locked in you are. A modular design can also make it easier to test and adopt improved models against saved business cases without rebuilding the entire solution.

Do you need AI specialists to run it?

Not a research team. Running a well-designed agentic process needs three capabilities: process and decision analysis, technical setup, and operating knowledge to monitor and adjust the agent. In practice that is a business owner who decides changes to the rules, an operator who reviews performance and escalations, and existing IT staff for infrastructure and security.

Rapidly growing and mid-sized companies often have valuable business expertise but limited time and specialized implementation capacity. MIT NANDA's 2025 research found that initiatives developed through external partnerships reached deployment approximately 67% of the time, compared with approximately 33% for internally developed solutions in its interview sample. These results do not prove that outside involvement causes greater success, but they suggest that specialized support can be valuable. The partnership should include knowledge transfer and training so the organization's team can operate and evolve the capability and choose whether to retain outside support.

Will AI agents replace your people?

For many rapidly growing and mid-sized companies, a valuable opportunity is increasing operational capacity without expanding staffing at the same rate. AI agents can take on certain routine judgments and repetitive activities, allowing employees to concentrate on complex exceptions, customer relationships, and work requiring experience. The effects on employment vary: some organizations redeploy staff, others reduce future hiring, and some adjust staffing levels. These outcomes depend on business strategy, implementation choices, and the work being automated.

How the initiative is explained matters. Employees who understand the purpose of the change and have a role in shaping it can provide essential process knowledge. Clear communication, meaningful involvement, and credible plans for training and oversight help organizations capture the expertise required to design dependable agentic processes.

How do you keep your experts' know-how when they retire?

A valuable asset in many rapidly growing and mid-sized companies is the judgment of experienced employees: how they price an unusual order, identify material exceptions, or understand a customer relationship. Much of this expertise is undocumented. The approaching retirement of experienced employees adds urgency. The Alliance for Lifetime Income identifies 2024 through 2027 as the Peak 65 period, when more than 4.1 million Americans are expected to turn 65 each year. Turning 65 does not necessarily mean retiring, but the demographic shift reinforces the importance of preserving critical knowledge.

Capturing expert judgment takes structured work with real cases rather than relying on a questionnaire. Review representative cases with experienced employees and record the evidence they considered, their decisions, and the reasons behind them, especially for exceptions. Documented decision rules can support employee training, auditability, and the implementation of AI agents. When this knowledge is maintained under clear access rights and governance, it can remain available after experienced employees leave and become a lasting organizational asset.

The real question: what will you own when it's done?

Price matters, but the larger question is what the investment leaves behind. Metered services can be the right choice for occasional or variable work. For steady, decision-intensive processes, a properly structured owned or hybrid capability may provide stronger long-term control: a working system,

documented decision logic, retained organizational know-how, and people capable of operating and improving it. The best choice is the one supported by the process economics, required safeguards, and the organization's operating needs.

Questions answered in this paper

Each question below has a short, direct answer in Agentic AI Costs and Ownership: 8 Key Questions Answered for Rapidly Growing and Mid-Sized Businesses.

- How much does this actually cost, and what's the ROI?
- What is a token, and what does it cost in real dollars?
- What does it actually cost to run an open-weight model ourselves?
- Are open-weight models really good enough? Aren't they a step behind the big AI companies?
- How do we avoid AI vendor lock-in?
- We don't have AI specialists. Can a company our size actually do this?
- Will this replace our people?
- How do we capture what our best people know before they retire or leave?

Sources

1. Gartner, "10 Best Practices for Optimizing Generative and Agentic AI Costs" (June 2026), as reported by Campus Technology. https://campustechnology.com/articles/2026/06/22/gartner-half-of-gen-ai-projects-could-exceed-budget-by-2028.aspx
2. Anthropic Engineering, "How we built our multi-agent research system" (June 2025). https://www.anthropic.com/engineering/multi-agent-research-system
3. Salesforce, "Salesforce Introduces New Flexible Agentforce Pricing" (May 15, 2025). https://www.salesforce.com/news/press-releases/2025/05/15/agentforce-flexible-pricing-news/
4. Stanford Institute for Human-Centered AI, AI Index Report 2025 (April 2025). https://hai.stanford.edu/ai-index/2025-ai-index-report
5. MIT NANDA, The GenAI Divide: State of AI in Business 2025 (2025), as reported by Fortune. https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/
6. Alliance for Lifetime Income, Peak 65 research, as reported by the Texas Hospital Association (2024). https://www.tha.org/blog/the-silver-wave-record-high-4-million-americans-will-turn-age-65-in-2024
7. Gartner, "Gartner Predicts AI Inference Costs Per Agentic Workflow Will Increase More Than Fivefold Through 2028" (August 17, 2026). https://www.gartner.com/en/newsroom/press-releases/2026-08-17-gartner-predicts-ai-inference-costs-per-agentic-workflow-will-increase-more-than-fivefold-through-2028
8. OECD, Benefits of AI Openness: OECD Discussion Paper for the G7 (May 29, 2026). https://doi.org/10.1787/746e8c9a-en
9. Salesforce, Agentforce Pricing (current product pricing and purchasing options). https://www.salesforce.com/agentforce/pricing/

About The Inteq Group

The Inteq Group helps organizations strengthen business performance through business process reengineering, business systems analysis, and agentic AI solutions built on open-weight technology deployed in infrastructure the client controls.

inteqgroup.com | 800-719-4627 | info@inteqgroup.com