Geordie

RESEARCH PAPER

Understanding the cost, the output, and the ownership gap across your agentic workforce

A guide for security, IT and finance leaders.

Based on insights from 50+ agent-forward organizations.

Foreword

Over the past three months, our team has spoken with more than fifty organizations running AI agents in production. Most of them found us through security, but the questions that kept coming up were economic ones: what is our agentic workforce actually costing us, what is it producing, and why can nobody in the room connect the two?

The enterprise is building a new labor ecosystem. Agents are researching, analyzing, building, and communicating across the same systems that people depend on every day. They are doing real work, and in many of these organizations, they are doing a growing share of it. The challenge is that the economics of this workforce remain almost entirely opaque. Most teams can see what they spent. Very few can say what that spend produced, whether the output was worth the cost, or which agents are earning their keep and which are burning tokens with nothing to show for it.

That gap kept surfacing because the instrumentation that observes agent behavior for security purposes captures the same data that explains agent economics. The record of what an agent did, which tools it called, and how it behaved is also the record of what that work cost, what it produced, and who should be accountable for both. Cost intelligence was a natural extension of what Geordie already understood about agents, and the conversations in this paper confirmed why it matters.

Behavior is the key commodity of what comes next. The organizations that understand what their agentic workforce is actually doing, and can connect that understanding to cost, output, productivity, and accountability, will be the ones confident enough to give their agents more responsibility, not less. Our mission is to make your agentic enterprise work, on your terms. This paper is what those conversations taught us, and it is honest about what we can answer today and what nobody, including us, has solved yet.

Henry Comfort
CEO & Co-founder, Geordie

Table of contents

ABOUT THIS RESEARCH

Insights drawn from conversations with agent-forward organizations running AI agents in production.

Organizations	Conversations	Org. size	Window	Geography
50+	**80+**	**750+**	**Jul–Sep '26**	**US & UK**

Industries

- Financial services
- Technology and SaaS
- Healthcare and biotech
- Private equity and alt. assets

Titles

- CISOs & Heads of Security
- VPs & Directors of IT/SecOps
- AI Strategy & Ops leads
- Finance leaders

IN BRIEF

Most organizations running AI agents can already see what they spent last month. Far fewer can say what that spend was bought. A billing dashboard totals tokens, seats, and API calls, yet it does not say which agent did the work, whether the work was useful, or whether the same outcome could have been reached for a tenth of the cost.

That gap is the reason agent spending has become the first material budget line in a decade to arrive without a natural owner. The reason the only lever most teams have found so far is a blunt one: stop spending, and work out the details later.

Three themes kept surfacing across these conversations, and this paper is organized around them. Each section covers what the conversations revealed and what the furthest-along organizations have done about it, closing with a synthesis of the maturity path and where the category is heading.

Ownership and accountability

Agent spend has landed across security, IT, and finance without attaching itself to any one of them.

Who owns it? | Who sees it? | Who can act on it?

Explainability and attribution

Billing dashboards show what was spent, not which agent drove the cost or whether the work was worth it.

Why doesn't the total explain itself? | What's missing between the bill and the work? | How much vs. why?

Response and control

When these gaps go unaddressed, the rational response is a freeze, and every leader here called that a failure of the tooling.

What are teams doing beyond freezing? | How do they want to intervene?

Ownership and accountability

Who owns the agent spend, who sees it, who can act on it.

Why doesn't anyone own this budget?

Cloud spend, when it arrived a decade ago, attached to an existing owner: IT's budget. Agent spend has found no equivalent home. Security can often see the activity (same instrumentation that watches for risk), while finance holds the budget and the authority to act. Very few have joined the two.

> [Ownership] is not clear. Finance still owns the IT and security budget, but it's a strange question: who decides, who signs off, and who decides what to buy.
>
> **TECHNOLOGY LEADER** · Enterprise customer

Some teams simply decline the ambiguity: one team owns agents' security posture, another owns what they cost. Nobody owns both, so nobody is positioned to connect them. Cloud spend had a decade to find its owner before agent spend landed on top of it; the tooling to attribute activity is only now catching up.

> We're at the beginning phases of trying to figure out what these tools do and then what it costs... I'm still trying to determine what the cost is so I can win that over with my CFO.
>
> **DIRECTOR OF ENTERPRISE SECURITY** · B2B software company

The scale of the gap often surfaces through individual outliers, not aggregate reports. One AI-native software company found a single employee's annual AI spend had reached over $100,000, with several others over $5,000 a month, unnoticed rather than hidden. High spending isn't necessarily bad; the problem is nobody can tell whether it's justified without first seeing it, then investigating the work behind it.

Who can see the cost data, and is it the right person?

A cost figure only works if it reaches somebody who can act on it. This is where more programs fail than on measurement. One organization pushed back on filing cost anomalies under security risk: a COO or finance lead needs to see what agents cost, not hold a security login or see threat intelligence alongside a spending question. If cost lives only inside a security tool, the people with budget authority can't see it.

Attributing spend to a named individual raises a question most AI-cost writing skips: the moment cost data attaches to a person, works councils, privacy officers, and European employment law all have a view. Several customers were already designing access tiers before being asked: a manager sees their team, an individual sees only their own usage, since retrofitting tiering is harder than designing it in from the start.

Asked what the output should look like, several leaders described a slide, not a dashboard: adoption trends, cost outliers, and a short answer to “is this sustainable” are what actually reach a board, and whoever owns that slide needs the cost data without a security login.

What does devolved ownership look like?

This is the groundwork for the top of the maturity curve. The most advanced organization in this set described wanting to devolve the budget entirely: hand it to the business unit that benefits from the agent's work. But that only works once the unit can already see its own consumption, recognize its own agents in the figures, and act without routing every decision through a central approver.

Explainability and attribution

Why the total doesn't explain itself, and what's missing between the bill and the work.

Why does the same waste keep showing up?

Across this research, seven waste patterns came up frequently enough and consistently enough across industries to constitute a taxonomy. They are listed here in order of how often they were cited, from most to least common.

Pattern	What it looks like	Whose problem
Oversized model	A frontier model drafting an email or checking the weather	User habit
Token maxing	Volume generated to look productive, producing little	Management incentive
Infinite loop	An agent repeating a task with no exit condition	Agent design
Inactive agent	Still connected, still billable, never invoked	Lifecycle
Low cache rate	Context re-sent that should have been cached	Architecture
Duplicate agents	The same agent rebuilt in four departments	Governance
Budget overrun	An agent burning far more than its intended allowance	Finance

Figure 1. Seven recurring waste patterns named across customer conversations, and who is best placed to act on each.

Not every organization is even in the consumption economy. A large share pay per seat for agent-enabled tools, and their waste has the same shape in a different currency: shelfware.

> People get a license, use it for a week, then it sits dormant for three months, while somebody else can't get a seat because we've used them all up.
>
> **IT LEADER** · Enterprise customer

The largest savings, though, rarely sit in architecture. They sit in habits nobody has been shown the cost of. Across the customer base, the single most concrete example was also the simplest: sending an agent a screenshot of a document, rather than pasting the same text, can cost several times as much for an identical outcome, because the agent has to do the extra work of reading the image before it can do anything with it.

Tie a KPI to token consumption, and you get Goodhart's law in a week

One of the more distinctive patterns in these conversations was also one of the most avoidable. More than one organization, looking for a simple way to measure AI adoption, reached for the only number readily to hand: tokens consumed. Staff, not unreasonably, responded to the incentive they were given, and used more tokens to look busy, whether or not the extra volume did anything useful. It is a textbook case of Goodhart's law: when a measure becomes a target, it ceases to be a good measure. Token count is a consumption metric wearing a productivity metric's clothes, and publishing it as a leaderboard corrupts it immediately. Any cost model worth building needs a denominator, work completed, or it will actively reward waste rather than discourage it.

That is the honest core of the problem, and it is worth stating three complications plainly, because a piece that skips them reads like a brochure.

- **Higher spend is sometimes the success signal.** An agent getting heavy use may simply be popular, doing more work for more people, with cost rising as a consequence of the value it creates rather than in spite of it.
- **Automatic model rerouting is oversold as the fix.** Once a task genuinely needs a given level of capability, the price gap between comparable models is often a couple of cents, and routing rarely moves the number that matters.
- **The harder half of the problem is still unsolved.** Most organizations can now say what an agent did and what it cost; very few can yet say, with confidence, what it actually achieved.

Where you instrument decides what you can price

Wherever the counting starts, it is bounded by where the organization can observe at all. A tool that lives at only one layer can price only what crosses that layer, and agents are not considerate enough to stay on one.

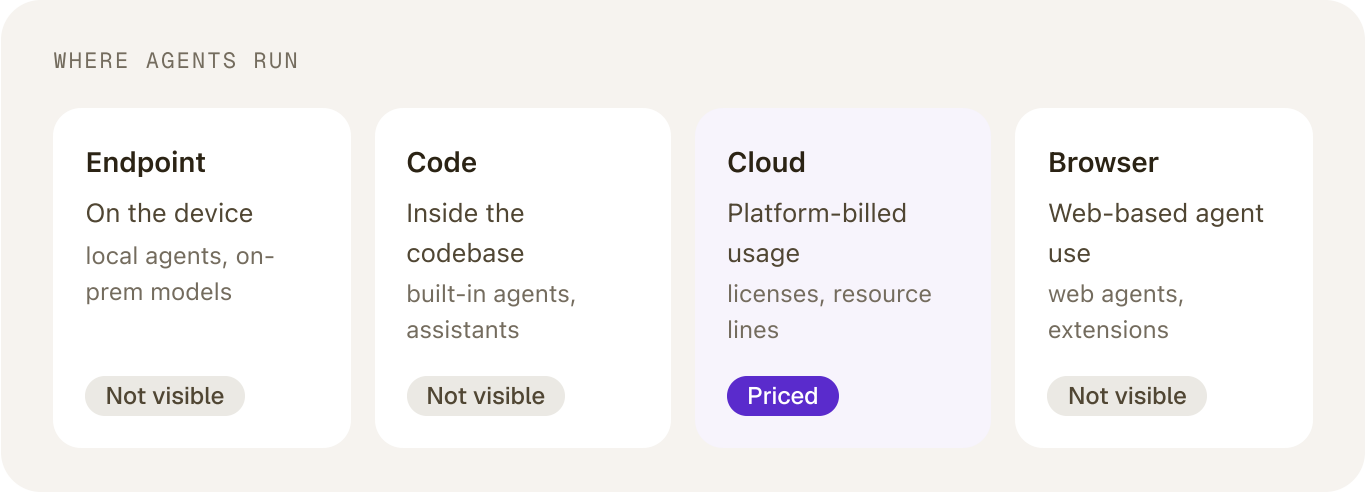

Figure 2. Where a cost tool observes activity decides which agent spend it can see and price. A provider's own billing boundary has a second blind spot: spend that moves to a locally run model never crosses it at all.

Cost pressure has also started to change where agents run at all. Several organizations described buying dedicated hardware to run smaller models locally, or routing cheap tasks to an on-premise model first and only reaching for a frontier model when the local one cannot cope. Each of these moves is invisible to any tool that watches spend only at a provider's own billing boundary. The meter a cost tool reads is not always the meter the money runs through.

> We can't get visibility unless we talk to our [AI vendor] rep... I'm on a seat-based plan and not necessarily token consumption-based.
>
> **SENIOR TECHNICAL LEADER** · Healthcare technology and services

The problem deepens when vendor introductory credits mask the real run rate entirely. One engineering leader at an enterprise software company described the pattern as a ticking time bomb.

> You've got 200 people that are using free included usage, but they're not showing up in the spend report... The reality is our budget per person is much less than the credit bucket that they've been given. So you're going to go from a complete blind spot to all of a sudden basically unaccounted or unforecasted spend.
>
> **ENGINEERING LEADER** · Enterprise software company

What does attribution actually require?

Every organization in this set that made progress started with a count rather than a policy, and the order matters more than it sounds. A policy written against the wrong denominator is a set of rules for a population that does not match the one you have. Counting is also the cheaper of the two, taking days rather than the weeks of committee time a policy consumes before anyone can act on it.

Two decisions shape what that count is worth. The first is what you agree to call an agent. A working minimum, a reasoning core with at least one tool connection, keeps copilots, scripted assistants, and multi-agent architectures in one list rather than three, and it is deliberately wide, because the agents that turn out to be expensive are rarely the ones that were formally adopted. The second is whether you count what is registered or what is running. A register records what somebody remembered to declare, on the day they declared it. Agents created by other agents,

tools added after the review, and instances still running for a project that ended two quarters ago are all invisible to it. Expect the running number to be higher, and in most cases materially higher: in one documented case a company found 327% more agents in production than its own technology and security teams had estimated, with no reconfiguration required to surface them.

What a first count should produce. Two numbers: the share of the estate under continuous discovery, and how spend is distributed across it. The first shows how much of the picture you actually have; the second shows where to look, and it is almost always more concentrated than people expect, which is what makes the next stage tractable.

Attribution is the stage where most programs stall, and rarely for want of effort. The word carries three separate claims: that a cost can be traced to a specific agent, to a specific action, and to a named owner. A tool can satisfy the first without the other two while still describing itself as attributing spend. Four things have to be true of the underlying record for that claim to hold:

- A stable agent identity that survives redeployment and renaming.
- An owner chain that extends through sub-agents.
- Action-level records, not session totals.
- A stated method for turning tokens into money, where estimation is legitimate and opacity is not.

There is a test that settles where an organization really sits, and it takes an afternoon rather than a procurement cycle. Pick one agent, take one day, and reconstruct what it did and what that cost, end to end. How long that takes, and how much of it has to be assembled by hand, is a more honest measure of maturity than any dashboard.

> Cost intelligence is not a way to spend less on agents. It is the precondition for spending safely.

Response and control

What teams are doing about it beyond freezing, and how they want to intervene.

Why is the freeze the default?

When a leader cannot separate good spend from wasteful spend, the rational response is to halt. That pattern recurs across conversations with organizations running agents at scale: a credit freeze, a blanket approval requirement, a moratorium on new use cases, each put in place because a finer instrument does not exist yet.

The freeze is a rational response to missing information, and it is also, on its own terms, a failure, because it slows down the exact activity the organization has usually been told, from the board down, to accelerate. The more consequential version of this story is not the team that overspent, it is the team that stopped. One technology leader described a specific decision to halt, taken after a single legitimate use case, run for three days, produced a cost nobody had predicted and nobody could explain.

> “
>
> We had an incident recently where spending got out of control and we didn't know it was happening. So you've already shown us something that would have helped us... stop that.
>
> **SENIOR TECHNICAL LEADER** · Healthcare technology and services

How should we budget agents?

Two customers, unprompted and independently, reached for the same analogy: an agent is closer to an employee than to a server. You would not add headcount without a job description, and you would not quietly expand someone's responsibilities without re-approving the change. The same discipline applies to an agent, where a widening scope of permissions is a decision rather than a side effect, and deserves the scrutiny a promotion would get.

> “
>
> You don't change a job description without getting it approved. It's similar with agents. You don't give one more responsibility without understanding the consequences of that.
>
> **TECHNOLOGY LEADER** · Real estate and development company

The primitive customers keep asking for by name, and cannot yet buy, follows directly from that framing: a declared budget per agent.

- **Set the number from observed spend**, not from a guess. Two weeks of measurement gives a better starting figure than any estimate.
- **State it in money**, per agent, per month, because token ceilings age badly as model prices move.
- **Revisit it when the agent changes**, not when the quarter does, because a new tool, a wider permission scope, or a change of model are all events that should reopen the number.

> “
>
> If we budgeted an agent at $1000 a month and it's burning $10,000, that's either getting a lot more use than we expected, or it's badly designed. Either way, we need to know.
>
> **TECHNOLOGY LEADER** · Real estate and development company

The same customers who raised the obvious objection, that estimated cost is never perfectly precise, dismissed it themselves once they said out loud what the number was for. Nobody needs $100 versus $102. Everyone needs to know when $100 becomes $1,000. The ratio is the product; the precision is a red herring.

Where should we sit on the intervention ladder?

Everyone in these conversations wanted the chance to coach an agent first; almost nobody wanted a hard stop as their only option. What we found instead were four levels of intervention, ranging from a gentle nudge to a full block, each taking progressively more control away from the agent and the person behind it.

1

Inform

Cost shown to the person who caused it, close enough to connect the two.

2

Nudge

A recommendation at the point of decision: the smaller model, the pasted text.

3

Shape

The cheaper path supplied by default, through the agent's own configuration.

4

Block

The action does not run.

> I'd love to enforce this... But I also need to have the confidence that I'm not nudging out people who are using valid accounts.

SECURITY / PLATFORM LEAD · Financial services firm

The question of when to block versus when to nudge came down, in several conversations, to whether the team understood the work well enough to know what was truly wasteful.

> Somebody looking for a recipe on Fable doesn't make a whole lot of sense for us to pay for that versus, maybe we're more willing to pay for it if they're looking at Haiku. So there could be use cases for us to block the use of specific models for specific use cases.
>
> **IT LEADER** · Enterprise SaaS company

Most of the waste in the taxonomy earlier in this paper responds to the first two levels, because most of it is habit rather than architecture. The patterns that genuinely need the fourth, a runaway loop with no exit condition, are also the ones that are easiest to specify precisely, which is convenient: the level that costs the most to get wrong is the one you need least often.

Two cautions are worth carrying into any design:

- **Blocking has a cost of its own.** It is not always smaller than the cost it prevents; more than one organization described downtime from a failed control as potentially more expensive than the incident it was there to prevent.
- **The control has to be consistent.** Whichever level you sit on, it has to reach the same decision every time. A control that behaves differently on Tuesday than it did on Monday is not a control, it is a second source of variance in a system that already has plenty.

The level belongs to the agent, not the organization: a nightly batch job with a predictable envelope and a customer-facing agent with a person waiting on it shouldn't carry the same setting. Choosing one posture for the whole estate repeats the freeze's mistake with better tooling.

Attribution is also what makes the freeze unnecessary. A credit limit applied to everybody is only needed while spend stays anonymous; once a cost has an owner, the conversation shifts from one blanket control to a set of specific decisions, each of which somebody is positioned to make.

What good looks like, and what to ask a vendor

We asked every organization in this set the same question: what does good look like for agent cost management? The most advanced among them did not describe a dashboard or a report. They described a state where business units own their own agent budgets, see their own consumption, and make their own decisions without routing every request through a central approver.

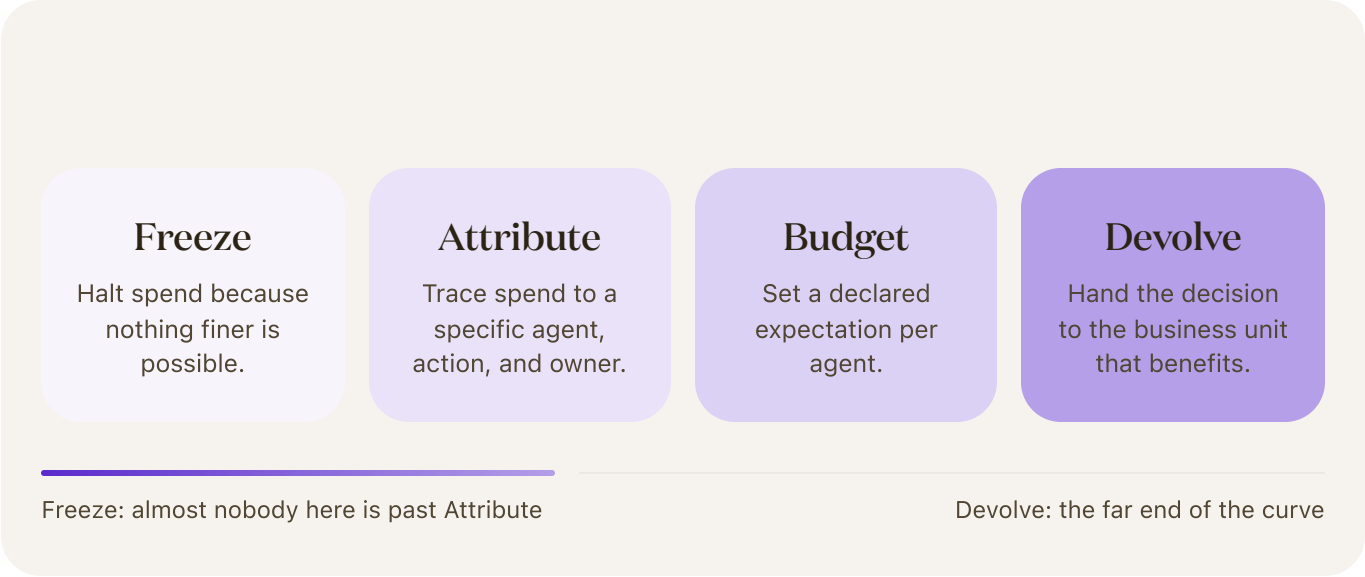

Figure 3. The four stages from an unexplained bill to a devolved budget, and where the conversations behind this paper actually sit.

The category itself is roughly two years old, and the maturity curve reflects that. Most of the organizations in these conversations are building attribution capability, the foundational stage that everything else depends on, and the natural starting point for a class of spend that did not exist at this scale until recently.

What the organizations furthest along had in common was sequence: each stage builds on the one before it. Attribution gives teams the data to set meaningful budgets, and budgets give business units the clarity to own their own decisions. Skipping a step, setting a budget before attribution is solid, or devolving a decision before the business unit can see its own consumption, creates the appearance of progress without the capability to sustain it.

What to ask a vendor

Most of what is sold today as agent cost management is a total, aggregated by platform or by team. The questions below came up most consistently across these conversations, in the order they build on each other.

- [ ] **Coverage that matches where agents actually run.** Does it reach endpoint, code, cloud, and browser activity, or only the portion of traffic that happens to cross a gateway?
- [ ] **Attribution to the agent, action, and owner.** Can the tool trace a figure back to a specific agent, a specific action, and a named owner, or does it stop at a number per team, per month?
- [ ] **A methodology it can show you.** Can the vendor explain how a cost figure is calculated, in enough detail that you could sanity-check it against your own provider invoices?
- [ ] **Data the right people can actually see.** Is cost locked inside a security tool the budget owner cannot log into, or is it structured for the audience that owns the spend?
- [ ] **Graduated intervention.** Can it inform and nudge before it blocks, and can a team choose where to sit for a given agent, rather than being offered only “allow” or “stop”?
- [ ] **A straight answer on what it cannot do.** What share of your estate does the tool expect to miss, and how would it know?

From tokens to outcomes

The token is not the ultimate unit of analysis for enterprises anymore. Work is. A token total does not say which agent consumed the tokens, what that agent was doing when it consumed them, whether the work succeeded, or where the instruction to perform it originated. The question that actually drives decisions is what work the consumption produced, and whether the cost was proportionate to the outcome.

The leaders closest to answering that question in these conversations were the ones measuring agent cost against the business result the agent was supposed to produce: ticket volumes, pull requests merged, claims processed, analyst hours displaced. When usage climbs and the business metric moves with it, the spend is earning its keep. When usage climbs and the metric stays flat, the agent is consuming resources without producing value.

Everything above is framed as a financial question, because that is usually how the conversation starts: a bill nobody can explain. In practice, the same instrumentation tends to surface two other exposures at the same time.

Security risk

A sudden, unexplained change in an agent's spend is frequently the first visible sign of something behaving abnormally: a compromised credential, a misconfigured integration, or an agent connected to a tool it should never have reached.

Operational risk

An agent stuck in a loop, or silently degrading in what it delivers while continuing to run, is both a cost problem and a reliability problem. The spend is the symptom that is easiest to measure.

The question worth putting to any vendor is not “can you show me what we spent,” but “can you also tell me whether it was safe and reliable, from the same underlying data?” Needing three separate tools and three separate logins means buying three partial answers instead of one connected one.

> **Geordie treats cost, security, and operational risk as three readings of the same underlying activity, because the instrumentation that observes an agent's behavior for security purposes is the same instrumentation that explains its cost.**

As cost pressure pushes more workloads toward on-premise and hybrid architectures, the share of agent activity invisible to gateway-based tooling will only grow. Cost intelligence that depends on being the chokepoint will miss the work that has moved off it, and that share is increasing.

It is worth ending on the limits rather than the capability. Connecting spend to what the work was actually worth is a different kind of problem. The organizations closest to it are still reasoning by proxy, using completion rates, rework, and time saved rather than anything that deserves the word measurement. And there is one question every leader in this set asked that nobody could yet answer: is my number normal? A peer benchmark, even an approximate one, is the single most requested capability this category has not yet built, and a fair signal of how early all of this still is.

This research draws on insights from more than 50 agent-forward organizations engaged by Geordie between July and September 2026. Findings reflect patterns observed across that set of conversations and are not a substitute for financial or security advice specific to any one organization's circumstances.

Geordie

geordie.ai