



Scale AI using your organization's intelligence with Microsoft IQ



03

Where enterprise AI stalls

06

The intelligence layer for every agent

07

Meet Microsoft IQ

08

One layer, four kinds of context

10

Microsoft IQ in action

11

Scaling agents you can trust

12

What shared intelligence delivers

14

Getting started



Where enterprise AI stalls

As enterprise AI adoption accelerates, organizations are deploying agents in growing numbers and planning to integrate many more in the coming years. Yet when they move from pilot to production, most agents stall — and usually for the same reason. An agent performs well in a proof of concept, then struggles the moment it needs to know how the business actually runs: who owns a decision, what the latest numbers say, or which policy applies.

Many organizations consider this a fragmentation problem: data, knowledge, and workflows are spread across many separate systems. But those symptoms trace back to a single cause: the organization's intelligence is not available to AI in a shared, usable form. **When AI can't reach that intelligence, the same recurring challenges appear:**

Agents lack true business context.

Agents built in silos miss what the data means, what matters in a given moment, and how people, systems, and information connect.

Context is rebuilt for every agent.

Each agent's grasp of the business is built by hand, tied to that one agent. The next team can't inherit it, so they start over from the same raw sources.

Costs escalate as agents scale.

One-off builds, duplicate integrations, and custom orchestration compound engineering and maintenance work with each new project.

Development is complex and brittle.

Custom plumbing and fragile logic rarely survive the move to the next scenario.

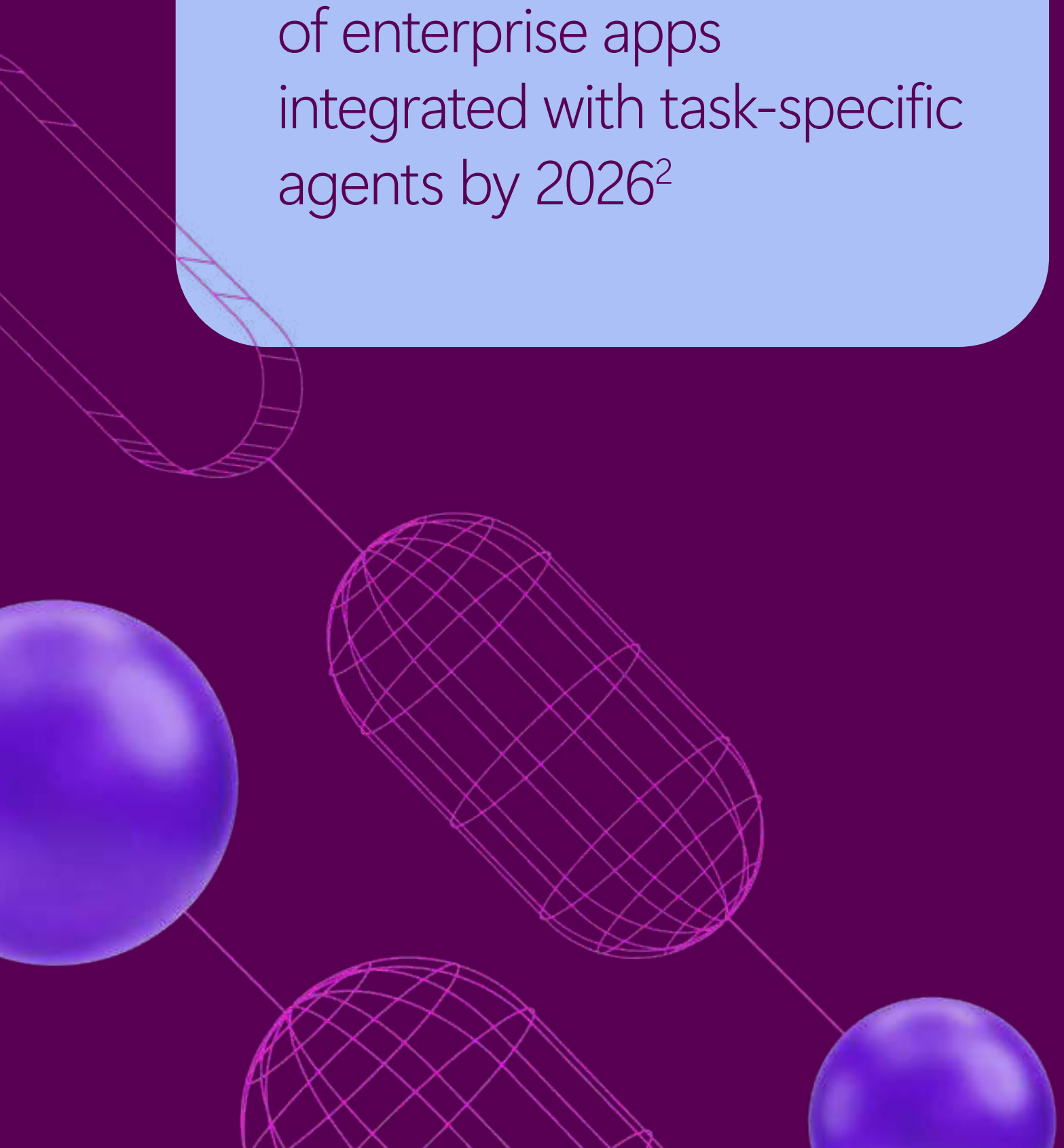
Security and governance are bolted on.

Without built-in visibility and control, leaders limit what agents are allowed to do.

Agents lack access to fresh, real-world knowledge.

Grounded in internal data alone, they can't draw on the continuously updated external information — market shifts, public developments, current conditions — that reliable, traceable answers depend on.





1.3B

AI agents by
2028¹

80%

of organizations intend to
integrate agents within one
to three years³

40%

of enterprise apps
integrated with task-specific
agents by 2026²

95%

of custom GenAI solutions
never reach production,
primarily due to brittle
workflows, lack of contextual
learning, and misalignment
with day-to-day operations⁴

1. Source: IDC FutureScape/IDC Info Snapshot research, 2025.

2. Source: Gartner, August 2025.

3. Source: Capgemini Research Institute. Harnessing the Value of Generative AI: Unlocking Scalable Advantage. July 2024.

4. Source: MIT NANDA, The GenAI Divide: State of AI in Business 2025, July 2025

The real barrier to scaling AI isn't access to capable models — it's that agents can't yet operate effectively across the enterprise.



Even as organizations have invested heavily in productivity tools, data platforms, and knowledge systems, their intelligence remains fragmented. The real barrier to scaling AI isn't access to models, agent count, or more tools and platforms. Capable models are already widely available, and waiting for a better one isn't always the solution, because the limit is how well agents can operate across the enterprise, not how smart the model is.

Doing that means giving agents a way to reach the intelligence already inside the organization: how its people work, what its business data shows, and what it has learned and documented over time. When that intelligence lives in a shared, reusable, trusted layer, every new agent inherits the organization's context on day one.

From there, the value of the intelligence layer compounds the more you use it. Each agent then adds what it learns back to the layer, so the next one starts from more than the last. Context accumulates instead of being rebuilt, and you stop paying the integration cost that used to climb with every new agent.

That layer is how organizations scale AI. Read on to understand:

- What the intelligence layer is and why it has become essential
- What Microsoft IQ is and the types of context it brings to every agent
- How it works in a real scenario
- How to take a first step using investments you already have

The intelligence layer for every agent

Think about what it takes for every capable employee to do their job well. They need to know four things: the people and workflows around them, the current state of the business, the knowledge the organization has already built, and what's changing in the world outside. Once an employee has all four, they can act with judgment and take on real responsibility.

An AI agent needs the same things to scale accurately. An agent that knows the people and workflows around it, sees the live state of the business, can reach the organization's accumulated knowledge, and stays current with the outside world can act effectively and be trusted with work that matters. An agent missing any of these four things produces plausible output that no one can trust enough to act on.

A unified intelligence layer is what supplies all of this. It's a shared, governed, continuously updated understanding of the organization that every agent inherits — the same understanding your best people operate from, made available to AI in a form it can use. A few years ago, a layer like this would have been an optional refinement. It has since become the foundation for AI at scale.

Why waiting has a cost

The necessity of a shared intelligence layer is acute and real. Leading AI models are now widely available and converging on what they can do, which shifts the lasting advantage in AI to the one thing each organization alone owns: its own context and data.

That advantage matters more today than ever before because agents are multiplying, and every new one needs that same organizational context to be effective and trustworthy. When agents are integrated one by one, each one is taught the business from scratch and a gap opens between how fast a company adopts AI and how ready it is to trust it.



Leading AI models are now widely available and converging on what they can do, which shifts the lasting advantage in AI to the one thing each organization alone owns: its own context and data.

That gap widens the longer it goes unaddressed. Going without a shared layer also carries a running cost. Every agent gets rebuilt and re-taught how the business works, which means months of custom integration and data re-plumbing per project. As adoption grows, duplicated effort and connector sprawl drive up costs. Results stay inconsistent from one agent to the next, so leaders hold AI back from high-stakes work and value never compounds.

Once you build that shared layer, scale starts working in your favor. Each new agent starts from one governed layer of context and adds to it, so the advantage compounds over time: every agent you deploy makes the next one faster to ship and the layer beneath them stronger. What used to cost more with each agent now returns more. **And because only you can build one from what your organization knows, it's the one part of an AI strategy a competitor can't copy.**

Meet Microsoft IQ

Your People. Your Data. Your IQ.

When your agents work from a single shared intelligence layer, each one starts already knowing how your organization works, and reasons from the same grounded understanding and the same trust layer as every other agent.

Microsoft IQ is that intelligence layer for the enterprise: a shared, continuously updated understanding of your organization that Copilot and every agent inherits by default. It takes the data, work signals, and knowledge already spread across your Microsoft estate and beyond and turns them into context that AI can use.

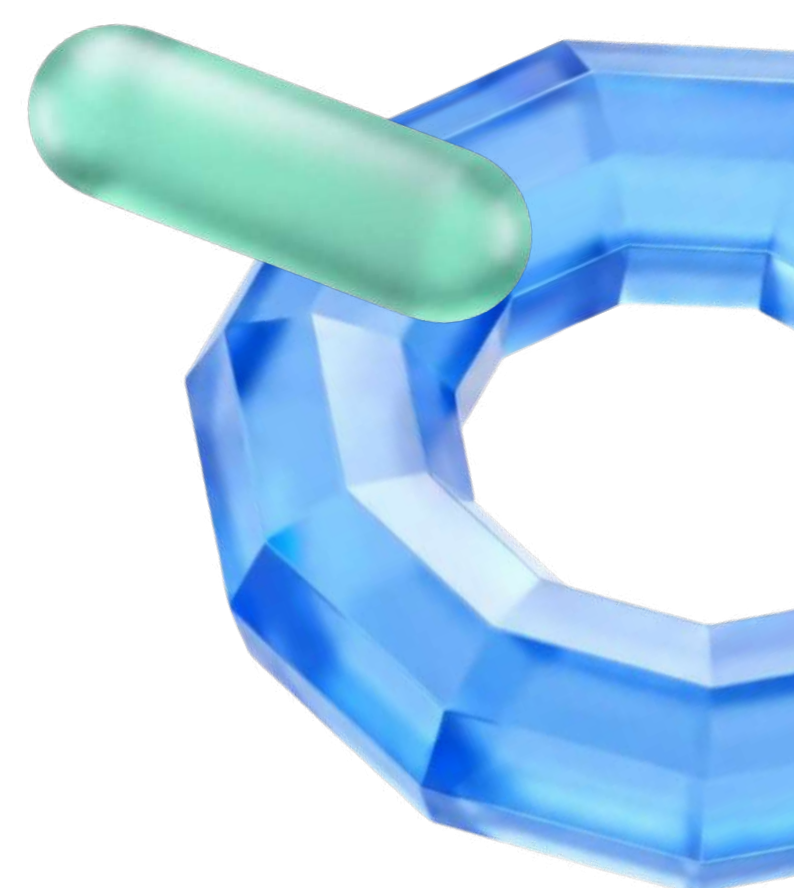
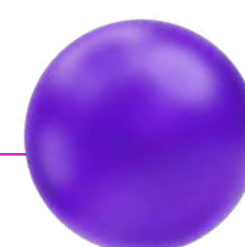
You activate it through Microsoft tools you already run, built on your own data and the investments already in place.

What your agents get from it is a single, governed understanding of the business to reason from. It spans how your people work across Microsoft 365, how your business is defined in your data and analytics, the knowledge held in your documents and systems of record, and the live information from the open web that keeps answers current.

That understanding is reusable and secure. Build an agent once and it knows your company on day one. Build the next and it reasons from the same grounded context, so each agent ships faster and the layer grows stronger as you add to it.

When you make Microsoft IQ your intelligence layer, it reaches into the workflows where work actually happens, across Microsoft 365, Dynamics 365, and the AI stack. Microsoft IQ runs on your organization's own data and intelligence, inside your tenant, with security, governance, and permissions enforced by default. Your data stays in your control.

And because Microsoft IQ draws on investments you've already made, your organization's context gets activated rather than rebuilt.



One layer, four kinds of context

Microsoft IQ brings together four kinds of context, each covering a different part of how your organization works. Together they give an agent the same rounded picture a capable colleague carries:

- Who does what and the workflows that connect them
- What the business data means
- What the organization has documented and decided
- What's happening in the world outside

Each kind of context is built on Microsoft tools and data already in place, so none of it has to be assembled from scratch.



Work IQ

How your people work

An agent that doesn't know who owns a decision or what was settled last week can't move work forward on its own. Work IQ gives it that footing — a real-time view of how the organization actually operates, built from the people, collaboration, and workflows already moving through Microsoft 365 and connected systems, and trimmed to what each person is permitted to see. In a typical Fortune 500, that view spans more than 600 TB of signals.

Ask an agent to move a project forward, and it knows who owns the next decision, what was agreed in last week's review, and which thread the open question is sitting in. And because the APIs behind [Work IQ](#) are built for the way agents work, an agent reaches all of it efficiently, using 80% fewer tokens and producing 2X more output per second than traditional approaches.

Fabric IQ

How your business operates

Across an organization, the same word can mean different things to different teams — revenue, customer, active account. A semantic layer gives your agents and your people one shared definition of the business, so they all interpret it the same way and analysis stays consistent no matter which agent produces it.

Fabric IQ turns raw data into that shared meaning. It can access more than 20 million semantic models already in use by 425K organizations, so your teams reuse the business logic they have defined rather than rebuilding it for each new project.

Shared semantics shorten the path from raw data to insight, and without context-rich grounding, AI results fall short.¹

¹Forrester, [Combine Semantics, Ontology, And Knowledge Graphs For AI-Ready Data](#)

Foundry IQ

How your agents unlock knowledge

When you ask an agent about a contract clause or an approval threshold, the output is only actionable if it's grounded in a governing document with a citation you can check, not a plausible guess. With a knowledge-retrieval layer in place, your agents respond from the organization's authoritative policies, documents, and knowledge bases, and return answers filtered by what each user is allowed to see.

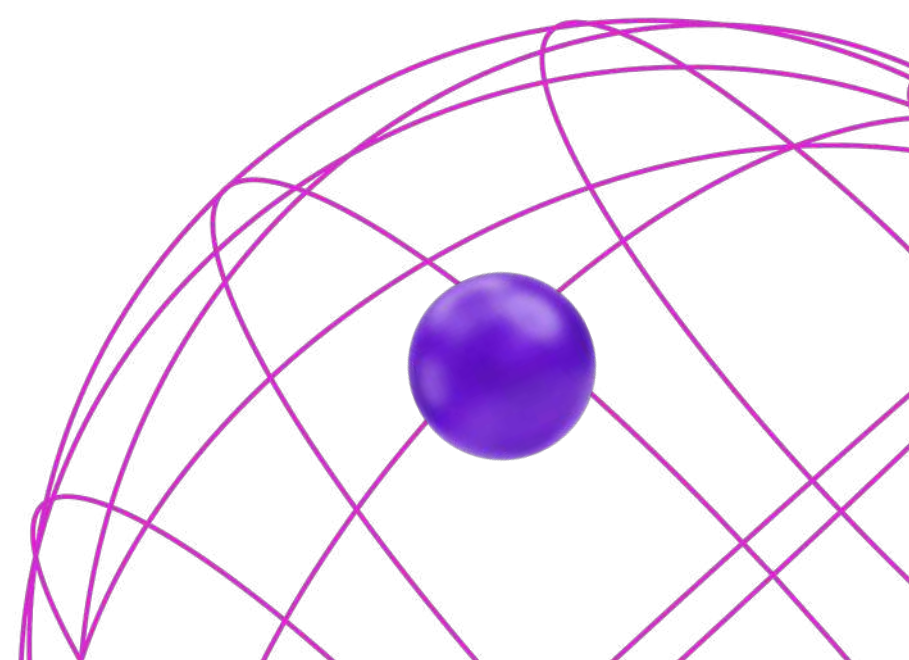
Foundry IQ provides that grounding. Its knowledge bases are reusable across agents, so you connect a body of knowledge once and every agent can draw on it. [Grounding agents in enterprise knowledge this way improves answer quality by 36%.](#)

Web IQ

How your agents quickly access the world

An agent working only from internal records can't see what changed in the world this morning. Web IQ lets your agents reason from current external reality, pulling in fresh, relevant information from the open web when a decision calls for it. Before an agent acts on a supplier risk or a market move, it can check what has been reported publicly in the hours since — information that postdates anything the model was trained on.

Web IQ provides that external grounding: web pages, news, and public signals, retrieved for agentic use, so answers stay current with information that changes after a model is trained. It runs nearly 2.5X faster than the next best alternative and draws fewer tokens per call for the same answer quality, so grounding in live web data keeps agents responsive in the middle of a task. It honors publisher preferences and runs through the same governed layer as everything else, so external grounding follows the same permission and trust rules.



Microsoft IQ in action

Microsoft IQ brings all four kinds of context together, so each agent inherits the whole picture — your people, your business, your knowledge, and the outside world — instead of a narrow slice stitched together for each project.

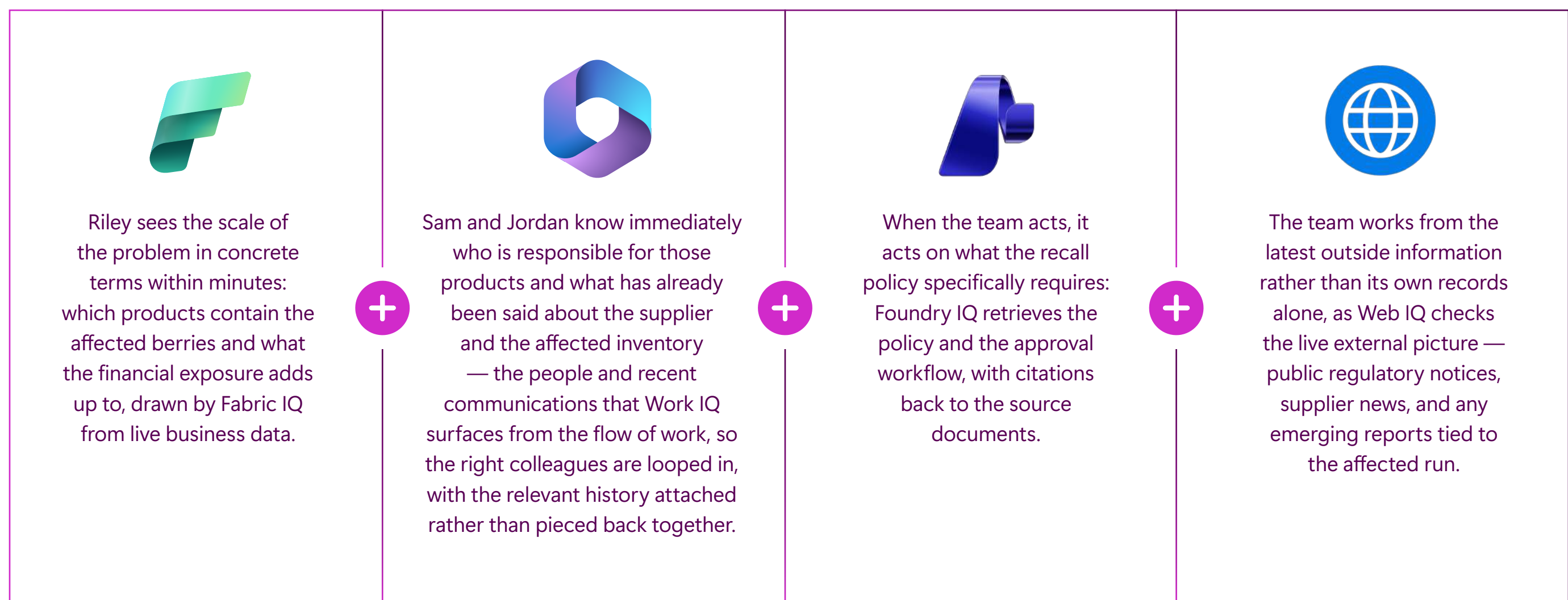
Consider how the four kinds of context work together when a situation demands a fast, coordinated response.

Sunrise Farms, a hypothetical food producer, learns of a possible contamination issue tied to a single production run of frozen berries. Responding effectively involves pulling together several parts of the business at once: which products are affected and what the financial exposure is, who needs to act and what has already been said, what the

recall policy requires, and what is being reported outside the company. Handled across separate systems by people alone, that coordination takes hours the situation may not allow.

At Sunrise Farms, the response to this contamination issue brings in several people — Jordan in purchasing, Sam in allocation, Alex managing a store, and Riley in financial planning — each working with agents inside their own tools, all drawing on the same shared intelligence layer.

The response begins with an agent monitoring supplier communications. The agent flags the contamination notice and checks it against the company's recall escalation criteria, confirming the issue is one that needs to be escalated. From there, each kind of context does its part:



What the team gets is one grounded picture rather than four separate answers. The buyer, the analyst, the store manager, and finance all move on the same set of facts. Every recommendation traces back to solid business data, the right people, cited policy, and the current external picture, so they can make decisions and take the next steps with confidence rather than stopping to reconcile four versions of the situation.

And the response leaves the organization better prepared for the future. What the agents established in this scenario stays in the intelligence layer: which products draw on which suppliers, who owns the response, which policy applies. The next time a supplier issue surfaces, the agents start from that groundwork rather than rebuilding it, and the response moves faster.

Scaling agents you can trust

Scaling agents only works if you can trust them with decisions your business depends on. That trust comes from the controls underneath every agent staying constant as you add more. The same governance, security, and permissions apply to all of them. What you rely on from one agent you can rely on from every agent that follows.

The traceability in the Sunrise Farms scenario holds everywhere Microsoft IQ operates, not just in a crisis. The policies governing one agent govern all of them, across Copilot, custom agents, and the applications your teams build. There is nothing to rebuild for each new use case, and all of it operates inside your tenant, so your data stays in your control. Intelligence moves to where your data already lives rather than the other way around.

Because that governance is consistent, you can account for what every agent does. Microsoft IQ grounds each recommendation and action in the shared context. Every one traces back to the business data, people, and policy behind it. Agents are monitored and managed centrally through Agent 365 and the Microsoft 365 admin center, which makes stepping in straightforward when you need to.

That oversight extends to security. An agent is no more of an entry point for attackers than the employee it works alongside: it carries a protected identity and is defended against threats through Microsoft Defender, at the same level of security that covers your people and infrastructure.

Security at the perimeter holds inside, too, where permissions decide what each agent can reach. Since context flows through Microsoft IQ, each agent reaches only what the person it acts for is cleared to reach. An agent working for someone in finance sees the financial records that person has permission to see, and nothing from HR they don't.

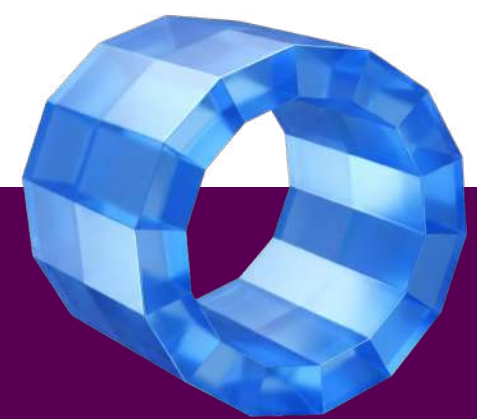
”

By integrating Foundry IQ, we provide a managed, permission-aware business context layer that connects marketing and brand knowledge into every agent so they can access the right information, at the right time, with the right governance.

Andrei Pop
Director of PM, Innovation
Sitecore

Sensitivity labels and data lineage travel with that information as the agent reasons across systems, enforced by Microsoft Entra and Microsoft Purview. The access rules you already maintain extend to every agent, with no separate model to manage.

Taken together, these controls add up to agents you can rely on as their numbers grow. Each one works inside rules you already enforce, every action it takes can be traced, and it reaches no further than the person it acts for. That is what makes an agent trustworthy enough to scale.

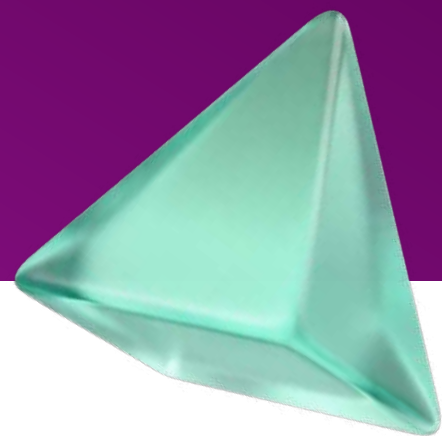


What shared intelligence delivers

”

We've been using Foundry IQ in our research and prototyping work, and the reusable knowledge base approach has cut a lot of the setup overhead we'd normally expect. Being able to ground agents in trusted enterprise content from day one, without rebuilding retrieval logic each time, has made early-stage experimentation noticeably faster and higher quality.

*Jane Chen
Lead AI Developer
Baringa Partners*



Once context is shared across every agent, the returns show up across everything your agents do: the decisions they make, the output they produce, the time it takes to get them running, and the trust you can place in them. And because every agent draws on the same layer, those returns compound over time: each new agent inherits what the last one learned, so your gains build with every agent you add rather than leveling off.

Decisions made with the full picture

The first return is broader context behind every decision. Agents reason with awareness of how the business actually works — your people, your data, your knowledge, and the outside world together — so their decisions account for the full picture rather than only part of it.

Accurate, grounded output

That context raises the accuracy of what agents produce. Output comes back grounded in reliable data and returned with citations you can check. You can trace the facts behind it instead of taking them on faith. An agent working from shared context answers from the record instead of filling gaps with plausible guesses.



Working agents in weeks, not months

Shared context also changes how fast your teams can move. Since each new agent starts with the organization's context already in place, your team avoids the integration and data re-plumbing that slows every project and the time from idea to a working agent drops from months to weeks.

Agents you can trust with high-stakes work

Quality output earns trust, and security and governance make it durable. Every agent works inside the permissions and policies you already enforce. Access never exceeds the person it acts for and every action traces back to its source. The more context and authoritative knowledge agents reason with, the more work you can trust them to handle.

Gains that build with every agent

The advantage compounds from there. Each new agent inherits the context built before it rather than starting from zero, so impact amplifies across the business with every agent your teams add.

Development and operational costs fall at the same time, as teams spend less on duplicated work and one-off maintenance. The same intelligence layer that made your first agent capable is what makes your hundredth faster, less costly, and better informed than the first.

Getting started

Building your organization's intelligence layer is less of a leap than it sounds — you've already done much of the work. Microsoft IQ builds on investments you've already made, so the first step is connecting the context that already lives across your tools, not standing up new infrastructure.

Start where your Microsoft estate is strongest. If your teams already work in Microsoft 365, model data in Power BI, or keep knowledge in SharePoint and connected systems, the raw material for your organization's intelligence layer is in place. From there, the work is putting that context to use for every agent you build.



Start scaling on your intelligence layer today.

Give every agent the context your organization already knows. Microsoft IQ is how you build the intelligence layer that makes AI work at scale.

- Explore [Microsoft IQ](#) and see how the intelligence layer comes together
- Read the [Microsoft IQ documentation](#) and the [Microsoft Agent Factory](#) white paper to go deeper
- Speed up cloud and AI adoption with expert support and funding through [Azure Accelerate](#)

©2026 Microsoft Corporation. All rights reserved. This document is provided "as-is." Information and views expressed in this document, including URL and other Internet website references, may change without notice. You bear the risk of using it. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.