ENTERPRISE AI AGENTS
A FIELD GUIDE

Bridging the Agent POC-to-Production Gap

How to get an enterprise AI agent from an impressive demo to a workflow your team can depend on.

tellius

Ajay Khanna
Founder&CEO, Tellius

JULY 2026

THE WHOLE GUIDE ON ONE PAGE

From POC to production, in one view

1 Why pilots die

No owner	The sponsor approves it, nobody answers for the output
Context	Right SQL, wrong definition of the number
No evidence	Unprovable, so a single wrong answer ends it
Last mile	The answer lands in a chat window, the work is elsewhere
Money	Cost scales first, with no budget to graduate into

2 Score yourself

Six dimensions	Use case · context · trust · workflow · governance · ownership
Four stages	Where your 0–120 score lands, and the constraint binding you there

3 What to build

Scope	One recurring decision, one owner, one baseline
Ground	Definitions, exceptions and hierarchies as versioned artifacts
Evaluate	A gold set on every change, plus three verification gates
Integrate	Arrive where the work already happens, then tier the actions
Operate	A named owner, monitored decay, a runbook
Compound	Use case two inherits the layer, not the lessons

4 What it’s worth

Weakest claim	Hours displaced, easy to compute but no cash moves
Strongest claim	External spend avoided, on a baseline set before go-live

WHY WE WROTE THIS

A note before you start

Hi — I'm Ajay.

I've spent the last several months in the room with commercial, finance, and market-access teams at some of the largest enterprises out there — with their analytics leaders, their platform teams, sometimes the consultants already on the payroll. Almost all of them had the same story: an agent demo that wowed a room, and then nothing. Six months on, it's a browser tab nobody opens.

The reasons rhyme, and they're almost never the model. It's the unglamorous part around it — the context nobody wrote down, the evidence nobody built, the answer that lands in a chat window instead of in the work itself.

So I wrote down what I've watched work and what I've watched fail: five failure modes you can catch in the first three weeks, an assessment you can run on your own program in about twenty minutes, and three real workflows taken apart end to end. Where someone published sharper evidence than mine, I've sent you there instead of paraphrasing it.

Ajay Khanna

Founder & CEO, Tellius

WHO THIS IS FOR

Whoever owns AI for the business — the head of the AI or innovation team, the analytics or data-platform lead, the digital exec who greenlit three agent POCs and has now been asked what came of them.

1

PART ONE

Why agent POCs don’t ship

Five failure modes that showupin the firstthreeweeksifyouknowwhatto look for, and turn nearly unfixable by month six.

THE STATE OF PLAY

Adoption is not the constraint, readiness is

35%

already use agentic AI, with 44% more planning to. MIT SMR & BCG

1 in 4

have not connected AI to their own systems and data. OpenAI

37%

use AI only at the surface, with little change to process. Deloitte

7×

more of leaders' AI work runs through purpose-built agents. OpenAI

The binding constraint is no longer the model or the tooling. It is organizational readiness, the same conclusion OpenAI reaches from a million business customers.

THE FIVE FAILURE MODES

Each one shows up in the first three weeks, and turns nearly unfixable by month six.

1 THE IMPRESSIVE DEMO WITH NO OWNER

An agent nobody is measured on is one nobody rescues when it breaks.

SIGNAL

The champion is whoever built it, and their title says "innovation."

FIX

Before you build, get the person measured on the outcome to commit out loud.

2 THE CONTEXT GAP

Flawless SQL, run against a definition of the number that isn't yours.

SIGNAL

Right query, wrong business answer.

FIX

Ask a question whose answer turns on a rule that lives nowhere in the schema.

WHY PILOTS DIE · CONTINUED

A working agent still dies without proof, delivery, or a budget

3 NO EVIDENCE, THEN THE TRUST COLLAPSE

One confident wrong answer, in front of the wrong person, ends the program.

SIGNAL

Accuracy gets described with an adjective, not a number.

FIX

Keep a gold set with known answers, and show lineage on every number.

4 THE LAST-MILE GAP

An answer in a chat window competes with the work, and loses.

SIGNAL

The agent lives at a URL you have to decide to visit.

FIX

Deliver into the deck, the queue, the inbox the work already runs on.

5 THE MONEY FAILURES

Priced as software it looks expensive. Priced against the labor it replaces, it is a rounding error.

SIGNAL

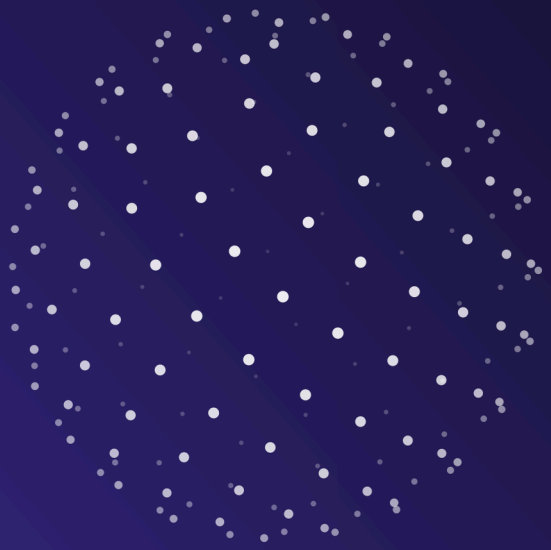
Nobody can state the cost of a single run.

FIX

Pick the budget you are replacing before go-live. It sets what you measure.

Notice what is missing from all five: the model. Each one is fixed by how you run the program, not by waiting for better AI.

2



PART TWO

The Agent Readiness Assessment

Thirtyquestions across sixdimensions—about twenty minutes to score your program out of 120 and find the one constraint holding it back.

THE ASSESSMENT · SCORE EACH 0–4

Score your program out of 120

0 1 2 3 4

0 not started 1 acknowledged 2 partial 3 works, one person true today. A scheduled initiative counts as zero. 4 systematic. Score what is

1 · Use case selection score /20

1.1	A recurring decision, describable in one sentence, that happens at least weekly.	0 1 2 3 4
1.2	Its current cost is measured in hours or cycle time, not estimated in the meeting.	0 1 2 3 4
1.3	A named business owner is measured on the outcome and has agreed to use it.	0 1 2 3 4
1.4	"Good enough" is a threshold agreed before the build, not after.	0 1 2 3 4
1.5	The workflow is bounded. We wrote down what the agent will refuse to do.	0 1 2 3 4

2 · Context & semantics score /20

2.1	Core metrics have single written definitions the agent reads at run time.	0 1 2 3 4
2.2	Entity hierarchies and their restatements are represented as-of.	0 1 2 3 4
2.3	Exceptions and carve-outs are encoded, not remembered.	0 1 2 3 4
2.4	Synonyms and local vocabulary map to the right objects.	0 1 2 3 4
2.5	Context is versioned and owned, with a process for changing a definition.	0 1 2 3 4

3 · Trust & evaluation score /20

3.1	A gold set of real questions with verified answers exists, and grows.	0 1 2 3 4
3.2	Every change runs against that set before release.	0 1 2 3 4
3.3	Answers carry visible lineage to sources, filters, and definitions.	0 1 2 3 4
3.4	The agent declines when it lacks enough to answer.	0 1 2 3 4
3.5	A correction path with an owner and an SLA, visible to whoever reported it.	0 1 2 3 4

THE ASSESSMENT · CONTINUED

The other three dimensions

4 · Workflow & actions

score /20

4.1	Output lands in a system the user already opens, not only a chat window.	0 1 2 3 4
4.2	The agent runs on a trigger, not only when a human remembers to ask.	0 1 2 3 4
4.3	It completes at least one action, not just an analysis.	0 1 2 3 4
4.4	Human-in-the-loop points are explicit and decided.	0 1 2 3 4
4.5	The workflow was mapped with the people who actually do it.	0 1 2 3 4

5 · Governance & security

score /20

5.1	The agent respects existing row and column-level permissions.	0 1 2 3 4
5.2	Every run is logged in a form an auditor would accept.	0 1 2 3 4
5.3	Residency, retention, and third-party terms are settled for the regions in scope.	0 1 2 3 4
5.4	Sensitive fields are handled by policy, not by prompt instruction.	0 1 2 3 4
5.5	Security, legal, and compliance reviewed the design before build.	0 1 2 3 4

6 · Ownership & economics

score /20

6.1	One named person owns the agent's output quality.	0 1 2 3 4
6.2	We know cost per run and at full rollout, with a routing plan for expensive paths.	0 1 2 3 4
6.3	We know which budget funds year two, and it is not an innovation fund.	0 1 2 3 4
6.4	Value is measured in the units that budget cares about, baselined before go-live.	0 1 2 3 4
6.5	An operating rhythm reviews usage, errors, and cost, with authority to act.	0 1 2 3 4

READ THE DIMENSION, NOT THE TOTAL

A 78 with a 6 in context is a rebuild. A 78 with a 6 in governance is a review cycle. Your lowest dimension sets the next quarter.

WHERE YOUR SCORE PUTS YOU

Four stages, and the constraint that binds each

Your total lands you in one of four stages. Each is held back by a different constraint, and that constraint is the thing to work on next.

Score	Stage	What is true	The binding constraint
0–40	Demo	The technology works. Nothing around it does. Usage is curiosity, driven by whoever built it.	Scope. You are building a capability, not automating a decision.
41–70	Pilot	Real users, a real workflow, and a trust conversation you keep having and keep losing.	Evidence. You cannot prove accuracy, so every expansion is re-negotiated.
71–95	Deployed	It runs and people depend on it, as somebody's second job, with cost now being noticed.	Operations and funding. The thing works and has no permanent home.
96–120	Compounding	Context, evaluation and delivery are shared infrastructure. New use cases take weeks, not quarters.	Demand. The constraint moves to choosing what not to build.

READ THE LOWEST DIMENSION, NOT THE TOTAL

A 78 with a 6 in context is a rebuild. A 78 with a 6 in governance is a review cycle. Your weakest dimension sets the next quarter.

WHERE TO START

Your next ninety days

The moves that clear the binding constraint, indexed to your stage. Between two stages, work the lower one, because the earlier constraint is always the one that binds.

Stage	Days 1–30	Days 31–60	Days 61–90
Demo	Stop building. Sit with one function until you can name a weekly decision in one sentence. Capture its baseline.	Get the owner's public commitment. Collect ten hard questions. Scope the refusal list.	Rebuild against that one workflow only. Nothing horizontal.
Pilot	Build the hundred-question gold set with the domain analyst. Get a number.	Move context out of prompts into a versioned layer. Add lineage and decline behavior.	Publish accuracy internally, weaknesses included. Model cost at 10×. Open governance review.
Deployed	Name a full-time owner. Write the runbook for failures you have already seen.	Instrument decay: accuracy over time, per-user usage, cost per run.	Take the value evidence to the budget line you intend to live in, before renewal.
Compounding	Publish an intake and prioritization process. Demand now exceeds c ap acit y.	Formalize stewardship of the context layer. Restrict who may change a definition.	Report in decisions and displaced work. Retire what no longer earns its cost.

3

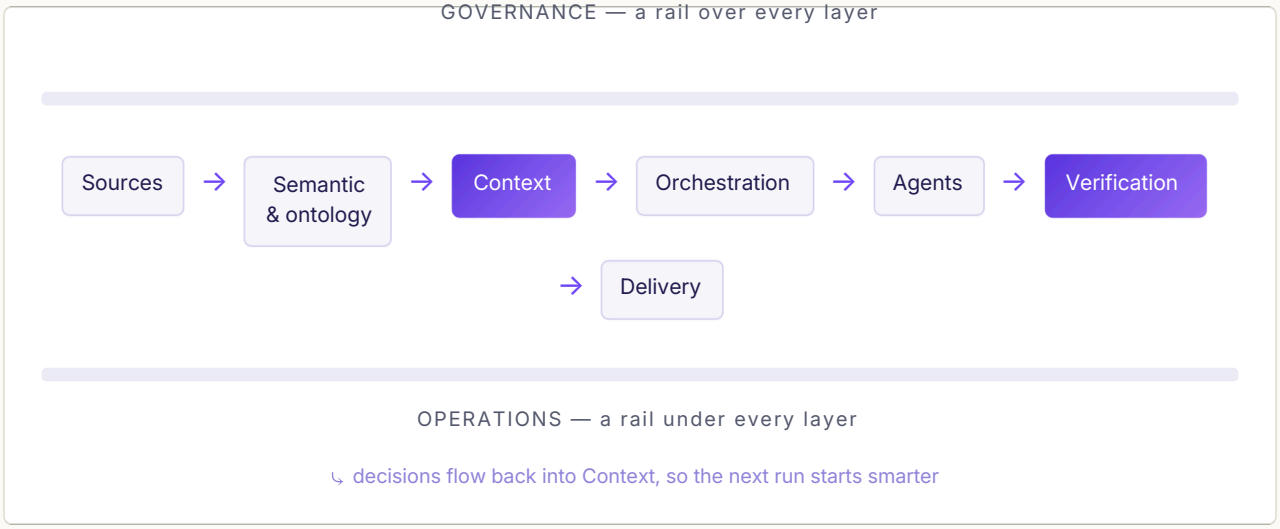
PART THREE

The production blueprint

The architecture, the context that feeds it, the checks that make it trustworthy, and how the output reaches the work. Build in that order — most failed programs did all of it eventually, in the wrong sequence.

THE BUILD, IN ONE VIEW

Every component, and who owns it



Context and verification (highlighted) are the two layers a demo skips and production cannot. Everything else here is available to your competitors on the same terms.

TWO LAYERS ARE THE DIFFERENTIATOR

Context and verification are what convert a fluent demo into an answer someone will sign.

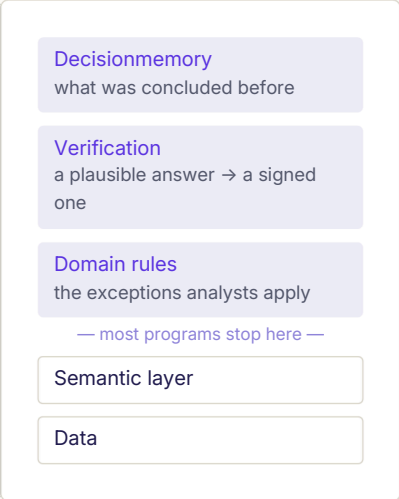
OWNERSHIP SPLITS FOUR WAYS

Governance and operations are constraints on every layer, not stages in the flow.

THE CONTEXT LAYER

Most programs build two layers and call it context

A semantic layer is necessary and inert. It tells an agent what a number is — which is why agents built on one are fluent and quietly wrong.



Three layers sit on top of the semantic floor. Almost every stalled program is missing them — not because they are hard, but because none improve the demo.

Layer	The question it answers	What breaks without it
Ontology	What exists, and how things relate — a territory contains accounts served by a rep, and a plan is governed by a contract.	The agent cannot traverse a question nobody scripted.
Semantic	How a number is computed — net revenue is gross less returns, excluding samples, by fiscal period.	Every team computes the same metric differently and defends its own version.
Domain rules	Which exceptions apply, and when — these account types are excluded, and the hierarchy restated in March.	Right query, wrong business answer — the most expensive failure to detect.
Procedural	How the work is actually done — to investigate a drop: rule out one-offs, then mix, then coverage.	The agent answers the question asked instead of the question meant.
Decision memory	What was concluded before — this pattern was an inventory movement, not demand, and who decided.	Every investigation restarts from zero, and nothing compounds.

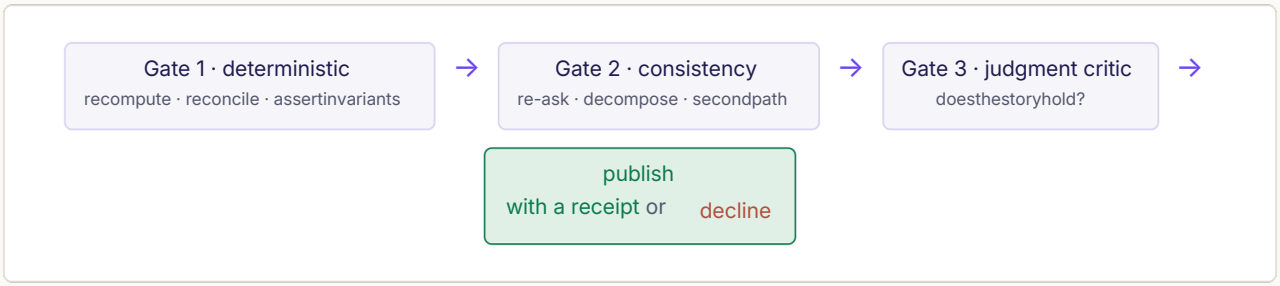
ONTOLOGY IS THE BACKBONE

A metric definition without an entity model is a formula with nowhere to go — which is why semantic-layer-only programs answer what they were built for and nothing adjacent.

WHAT MAKES AN ANSWER SIGNABLE

Verification, and the receipt underneath

Asking a model to check a model feels like verification and often is not — the checker inherits the blind spots of what it checks. Order the gates so the cheap, deterministic ones run first.



VERSIONED CONTEXT ARTIFACT
metric: net_revenue

formula: gross − returns
exclude: [samples, intercompany]
calendar: fiscal (ends Nov)
hierarchy: territory@2026-03
owner: fpna.commercial
version: 4 · effective 2026-05-12

RECEIPT ATTACHED TO THE ANSWER
sources: [snowflake, iqvia]
rows_in: 214,338 · rows_out: 6
gates: det ✓ · consistency ✓ · critic ✓
defs: net_revenue v4
diff: −6.0% vs prior mo
why: tier-2 access loss, 3 zips

THE LAST LINE IS THE ONE THAT SAVES DEPLOYMENTS

"Why is this different from last week?" has an answer — and the answer is not "the AI changed its mind."

WHAT ACTUALLY GOES WRONG

Eight ways an agent answer fails

Eight failure shapes, each with the check that catches it. A program that ships has a named defense for every row.

Failure	What the user sees	Why it is dangerous	What catches it
Silent filter	A confident number built on a subset, one region missing or one channel excluded.	The query succeeded and nothing looks wrong, so nobody looks.	Row counts in the receipt, reconciled to a known total.
Fan-out	Revenue inflated by a join that multiplied rows.	Large and directionally consistent, so it reads as good news.	Invariant assertions. Totals must tie to the source of truth.
Re-aggregation	An average of averages, or a rate summed across segments.	Plausible magnitude, wrong answer. It survives eyeballing.	Deterministic recompute from the base grain.
Calendar / period	Fiscal treated as calendar, or a restated hierarchy applied to a prior year.	Every comparison in the answer inherits the error.	Definitions resolved in the receipt, with as-of context in the layer.
No lineage	A number with no working shown.	Nobody defends a number they cannot trace, so they stop repeating it.	A receipt attached to every answer, one click away.
Run-to-run drift	The same question, twice, two different answers.	It destroys trust faster than being wrong. Unstable cannot be relied on.	A consistency gate, with cached, reproducible results.
Path inconsistency	The dashboard says one thing, the agent another.	Two truths, and no way to adjudicate between them.	Shared definitions, reconciled against the certified report.
Silent ambiguity	“Region” resolved to one of three hierarchies without saying which.	Defensible, and not what was asked. Found weeks later.	Restated interpretation up front, asking when ambiguity is material.

ARRIVAL BEATS ACCESS

Delivery, and routing the work

Two plumbing decisions that decide how much time the workflow actually saves: where the output lands, and which path answers each request.

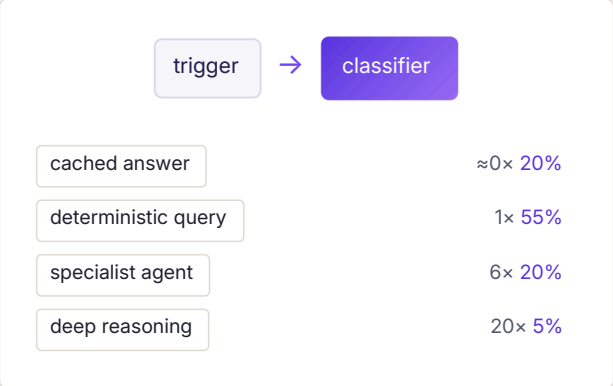
Tier the actions

Analysis that arrives — in the review deck, the queue, the channel, on the cadence the work already runs to — gets used by default. Analysis that must be fetched gets used by enthusiasts. Then tier what the agent may complete on its own, by reversibility and blast radius.

Auto	Reversible, internal, routine. Runs unattended, and logged.
Notify	Runs, then tells a human what it did, with an undo.
Approve	Drafts and waits. External, financial or ir reversible.
Refuse	Outside scope. Declines and names the boundar y.

Most workflows land near 70 / 20 / 10 across the first three tiers. If yours is 0 / 0 / 100, you have automated the analysis and kept the labor.

Route to the cheapest path that can answer



Escalate, never default: the deep path is what a failed check triggers, not where every question starts. Cap it at one escalation.

OPERATIONS

What breaks in month three

Nothing dramatic breaks. That is what makes this stage dangerous — accuracy degrades quietly for four mundane reasons.

DEFINITIONS DRIFT
Finance changes an exclusion, and the context layer never hears about it, and the agent reports last quarter's rule.

QUESTIONS GET HARDER
Users move from the questions you designed for to the ones they actually have. The gold set no longer represents the workload.

UPSTREAM DATA CHANGES SHAPE
A source adds a field, restates a hierarchy, or starts populating a column that was always null.

THE OWNER GETS PROMOTED
The most common operational failure. A runbook in one person's head is a resignation risk, not a technology risk.

MONITOR WEEKLY
Gold-set accuracy · usage by user, not in total · declined and failed runs · cost per run and total spend · median time to fix a reported error.

REVIEW MONTHLY
New failure classes · context changes shipped · questions the agent should now handle · agents that should be retired.

The tell: flat total usage hiding a shrinking group of heavier users. Aggregate numbers conceal churn. Per-user numbers do not.

SEVEN TESTS

Anatomy of a use case that works

Successful deployments look far more alike than the organizations that built them. A use case that fails two or more of these will consume a year and end in a demo.

- [ ] **It is a recurring decision, not a question.**
  Weekly at minimum. Frequency is what converts saved effort into money.
- [ ] **The trigger is not a person.**
  An event or a schedule starts it. Anything that waits on a human remembering caps out at the enthusiasts.
- [ ] **The input space is bounded.**
  Thirty question shapes you can ground and test, not “anything about the commercial business.”
- [ ] **The output is verifiable.**
  There is a right answer, or one that reconciles to something certified. Where correctness is unknowable, trust never accumulates.
- [ ] **It produces the first draft of the work.**
  A brief, a ranked queue, a filled template that is deliverable-shaped, not an input to the work.
- [ ] **Failure is recoverable.**
  One person’s objective visibly moves when it works, and irreversible actions sit behind a human gate.
- [ ] **The shape repeats.**
  The same mission applies across territories, brands or entities, so scaling is a copy, not a rebuild.

THE LAST TEST IS THE COMMERCIAL ONE

The first six decide whether it works. The seventh decides whether it was worth doing.

WHAT THIS LOOKS LIKE IN PRACTICE

Six jobs worth handing an agent first

Across the functions we work in, the same handful of recurring jobs come up first: frequent enough to be worth automating, and bounded enough to trust. Each runs as a mission that fires on a trigger, assembles the work, and lands where the team already looks.

Function	Mission	What it assembles	Where it lands
Pharma commercial	Weekly brand review	Performance vs. plan, the three drivers behind the movement, and the exceptions worth discussing.	The review deck, before the meeting rather than during it.
Field effectiveness	Target reach & quality	Which targeted prescribers are reached, which high-potential ones are missed, where activity is not changing behavior.	Field digest, with the call list updated.
Market access	Pull-through gaps	Where coverage exists on paper but scripts are not converting, ranked by recoverable volume, plan by plan.	Access team queue, assigned by territory.
Market access	Contract & coverage risk	Exposed volume when a payer position changes, the accounts affected, and a first-draft response.	Account brief for the market access lead.
Finance / FP&A	Variance root cause	Variance decomposed into price, volume, mix and timing, with the three largest contributors named and evidenced.	The review pack at close.
Procurement / S2P	Three-way match exceptions	PO, receipt and invoice mismatches ranked by value, with likely cause and a drafted supplier query.	Sourcing queue, assigned to a category owner.

WORKED MISSION · 1 OF 3 · PHARMA COMMERCIAL

The weekly brand review

A brand or analytics team pulls performance, builds the deck, and chases the "why" — one to two days per brand, every week. The meeting then re-derives numbers everyone already had.

1. Trigger — Monday morning, once the weekly sales and script refresh has landed.
2. Ground — Applies the brand's own definitions: prescription basis, sample and channel exclusions, fiscal calendar, and the territory hierarchy with its restatement dates.
3. Reason — Movement against plan and prior period, decomposed by geography, channel and segment, then isolates the three largest contributors and tests each against known one-offs.
4. Verify — Recomputes against the certified weekly report and reconciles totals. Where the certified source disagrees, it says so rather than choosing.
5. Deliver — Writes headline, drivers, exceptions and watch list into the existing deck template.
6. Remember — Retains what was flagged and concluded, so next week opens on the unresolved items.

CONTEXT IT MUST HAVE

Prescription definitions and exclusions · territory hierarchy with restatement dates · plan versions · a calendar of known one-off events.

WHAT STAYS HUMAN

The narrative judgment, and anything that reads as a recommendation to change promotional or clinical strategy. The deck arrives as a draft with the working shown.

THE WEEKLY BRAND REVIEW · WHAT ARRIVES

The output, and the receipt underneath

Illustrative output. Figures are synthetic — the shape is the point: a claim, its size, its cause, and whether it is real.

WEEKLY BRAND REVIEW · WEEK 30 · DELIVERED 07:04 MONDAY

Volume is down 2.1% week on week. Year-to-date variance to plan remains inside tolerance at −0.8%.

Three drivers. One region accounts for 61% of the decline, concentrated in two territories restated on 1 March — comparison is like-for-like. A second region is flat but masking a channel shift. The third contributor is a holiday week in the prior comparison period, not a real movement.

Two exceptions. One territory has been outside its control band for three consecutive weeks. One account shows a step change consistent with a formulary event rather than field activity.

Watch list. Two items carried forward from last week, one now resolved.

RECEIPT

interpretation: NBRx, wk30 vs wk29 defs: net_volume v4 (excl samples) sources: [snowflake, iqvia, veeva] rows: 214,338 → 6 drivers
gates: det ✓ · consist ✓ · critic ✓

disputed: certified report shows −2.3%, flagged not resolved

The disputed line is the one that earns trust. It would have been easier to pick a number.

WORKED MISSION · 2 OF 3 · FINANCE & FP&A

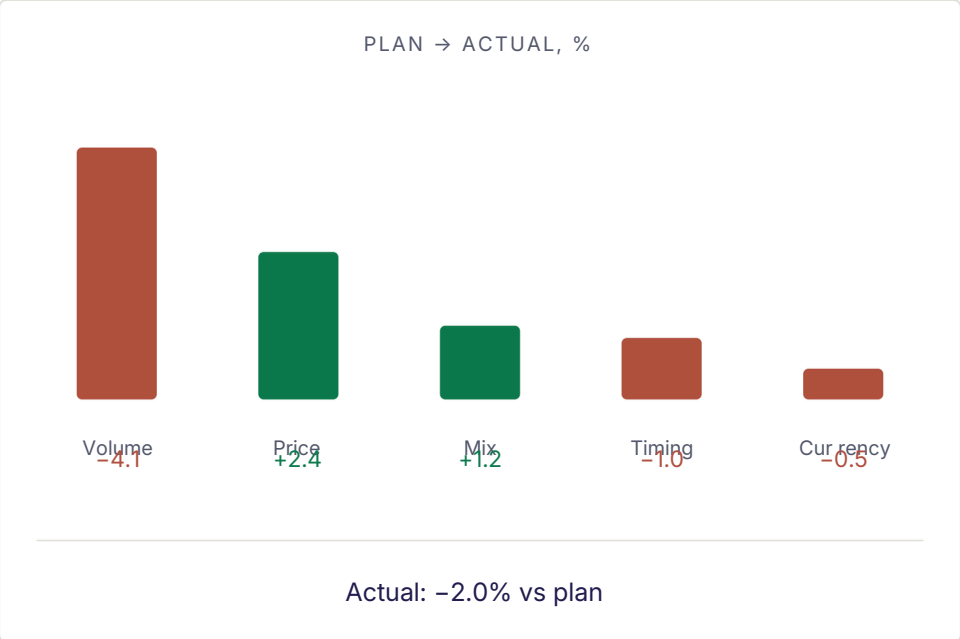
Net revenue variance root cause

Analysts spend the first week of close assembling the variance pack. Commentary gets written from memory, and two people produce different numbers for the same line because they filtered differently.

1. Trigger — The ledger close event, or the month-end flag.
2. Ground — Revenue definitions by entity, the currency policy and rate source, intercompany eliminations, and which plan version is in force.
3. Reason — Decomposes variance into price, volume, mix, timing and currency, ranks the contributors, and walks each down to the accounts behind it.
4. Verify — Components must sum to the reported variance, a hard invariant. Recomputes from base grain and reconciles to the certified statement.
5. Deliver — Drafts flux commentary per line into the close template, supporting detail attached to each claim.
6. Remember — Holds last period's explanations, so a recurring driver is labelled recurring, not rediscovered.

HOW YOU KNOW IT WORKS

The invariant does the work: if the decomposition does not tie to the reported variance, the run declines.



Volume is the whole story, partly offset by price. The bridge is both the deliverable and the test: the components must sum, or the run declines.

WORKED MISSION · 3 OF 3 · PROCUREMENT & S2P

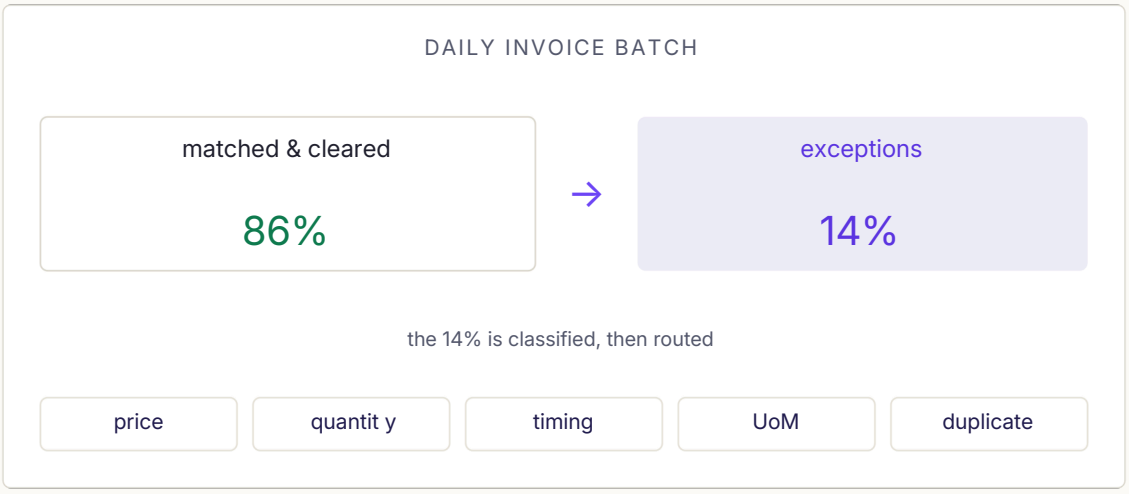
Three-way match & price leakage

The exception queue is sampled rather than cleared, because clearing it fully would take more analysts than the function has. Leakage surfaces months later in audit, when recovery is hardest.

1. Trigger — The daily invoice batch.
2. Ground — Contracted price lists with effective dates, tolerance rules by category, unit-of-measure conversions, and the supplier hierarchy.
3. Reason — Matches PO to receipt to invoice, classifies each break by price, quantity, UoM, timing or duplicate, quantifies recoverable value, and groups by supplier.
4. Verify — Recomputes against contract terms, checks duplicates against paid history, reconciles to the payables ledger before anything is raised.
5. Deliver — A ranked queue with a drafted supplier query per exception. Within tolerance it clears unattended. Above threshold it holds for a human.
6. Remember — Each query outcome feeds back, recovered or rejected and why, which makes next quarter's ranking better than this one.

WHAT STAYS HUMAN

Anything that touches the supplier relationship or authorises payment. Recovery claims are drafted, never sent.



An unclassified queue is a list of complaints. A classified one routes itself to the person who can close it.

HOW IT ACTUALLY HAPPENS

The first twelve weeks of a deployment that ships

Weeks one to six look like no progress — nothing is demoable until week seven, which is exactly why most programs skip to building and pay for it in week eleven.

Weeks	What happens	What slips here
1–2	Choose the decision. Sit with the function until the workflow is one sentence. Baseline captured, owner committed.	The owner will not commit — and the right response is to change the use case, not the owner.
3–4	Ground it. Definitions, exclusions, hierarchies and restatements into a versioned artifact.	Two teams disagree on a definition and nobody is allowed to decide.
5–6	Build the gold set. A hundred real questions, answers verified by the domain analyst.	The analyst has a day job. The step most often skipped — and skipping it is why week eleven fails.
7–8	Shadow run. The agent runs beside the humans, and every disagreement is logged and classified.	Calling it done here. Shadow running is where the silent failures surface.
9–10	Pilot with five users. Output lands in the real surface, approval tiers on, and corrections feed the gold set weekly.	Usage measured in aggregate hides four of the five drifting away.
11–12	Trigger it. Schedule or event fires it, owner named, monitoring on accuracy, cost and declines.	No budget line — that conversation starts in week six, not week twelve.

SHADOW RUNNING IS THE STEP TO PROTECT

Two weeks of the agent working beside people, every disagreement logged, buys more trust than any accuracy number — and surfaces the silent failures a gold set alone will not.

CASE EVIDENCE

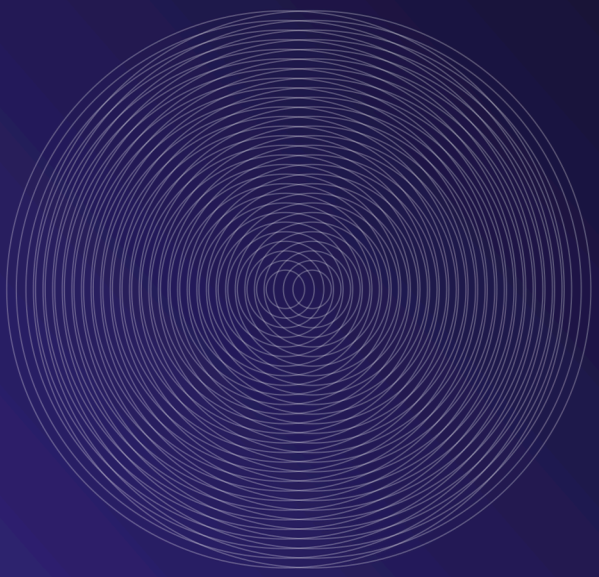
What one team measured, and what it confirms

Anthropic's data science and data engineering team published how they run self-service analytics internally. Their numbers are the closest public evidence to the argument in this guide — and several of them are uncomfortable.

What they report	What it confirms
Without procedural skills, accuracy did not exceed 21% on their evals. With them, consistently above 95%.	Procedural knowledge — the order an analyst works in — is the single largest accuracy lever, and the layer most teams skip.
Offline accuracy drifted from 95% at launch to 65% within a month, until maintenance was treated as engineering.	Decay is the default. A program without a maintenance loop loses a third of its accuracy in weeks, quietly.
Direct access to thousands of prior queries moved accuracy by less than a point. The answers were present and unused.	The bottleneck is structure, not access. Pointing an agent at more material is not grounding it.
LLM-generated metric definitions tested worse than a smaller, human-curated set.	The context layer cannot be generated. Ownership is the work, and it is the part that cannot be automated away.
An adversarial review step added 6% accuracy for 32% more tokens and 72% more latency.	Verification has a measurable price. This is why gates are tiered rather than applied to everything.
Every answer carries a footer showing source tier, data freshness, and who owns the model.	The receipt, in production — one of the few mitigations for silent failure, which they do not claim to have solved.

Source: Anthropic, "How Anthropic enables self-service data analytics with Claude," June 2026. Figures as reported by their team. The mapping to this guide is ours.

4



PART FOUR

Keeping it funded past year 1

Aworking agent still getscancelled ifthe wrong teamowns itornobodycan prove what it is worth. This part is about making it survive that.

OWNERSHIP

Why a center of excellence is usually the wrong owner

A center of excellence is good at starting things and structurally bad at owning them. It is funded per initiative, staffed by people whose careers advance by launching rather than operating, and organizationally distant from the consequences of a wrong answer. None of that is a criticism of the people in it.

Business function owns output quality and adoption — they carry the consequence.

Platform team owns the context layer, evaluation and cost — shared assets that degrade when owned locally.

The center keeps standards, intake, pattern reuse — and stops four teams building the same agent.

Agent programs lose budget arguments they should win because of the category they are filed under. Compared with a BI license, a production agent program is expensive. Compared with the analyst hours, contractor invoices and consulting SOWs it displaces, it is usually not close. The comparison has to be chosen before go-live, because it determines what you measure — and that evidence cannot be reconstructed afterwards.

Make the case against the outside spend it replaces: the work of a consulting engagement, done at software cost, with the before-and-after to show it. That lands with finance in a way "it improves analyst productivity" never will.

THE VALUE CASE

What it is worth, and what survives finance

Five ways to count the value, ordered by how well each survives a finance review. The weakest moves no cash. The strongest is the one unambiguously cash line.

Source of value	What to capture	Why it holds up
Capacity reallocated	Hours moved off rote work, and specifically what the team did with them instead.	The real version of "savings." Counting hours alone moves no cash and puts the owner on the defensive.
Throughput	Cases, reviews or decisions handled per person per period, before and after.	Productivity, not headcount. Nobody loses a job for this to move, so the business owner will help you measure it.
Coverage	Cases reviewed rather than sampled, and what the extra coverage found.	Value that was previously impossible, not merely cheaper. The most under-sold line.
Cycle time	Request-to-decision on the same workflow, before and after.	Week one instead of week four changes what is still recoverable.
External spend avoided	Statements of work, contractors, ad-hoc requests not raised.	The only unambiguously cash line. Strongest in a budget conversation.

TWO DIFFERENT ROOMS

Accuracy earns users and the owner's trust — lead with it there. Finance funds outcomes, so bring accuracy as proof the outcome is real, not as the headline.

THE COMPARISON THAT LANDS IS YOUR OWN

Peer benchmarks show a gap exists. You win the budget on your own before-and-after.

SCORE IT WITH YOUR TEAM

The readiness scorecard

Have the builder, the intended user and the budget holder score it separately. Where they disagree by two or more points, that gap is the finding.

Dimension	Q1	Q2	Q3	Q4	Q5	Total /20
1 · Use case selection						
2 · Context & semantics						
3 · Trust & evaluation						
4 · Workflow & actions						
5 · Governance & security						
6 · Ownership & economics						

Total / 120: _____ Stage: Demo (0–40) · Pilot (41–70) · Deployed (71–95) · Compounding (96–120)

SCORE IT WITH THREE PEOPLE

Whoever built the POC, whoever is meant to use it, and whoever pays for it. Where they disagree by two or more points on a question, that gap is the finding.

WHAT YOU WALK AWAY WITH

Your score across the six dimensions, the constraint binding you at your stage, and the next three moves worth making.

WHAT TELLIUS IS

This guide is the playbook. Tellius is how you run it.

Tellius gives everyone anAI worker thatshows upwith the work done. It reasons across your data with your business's context applied, produces the finished brief, pack, or play, and ships it into the tools you already use, deterministic and governed, so the number holds up in front of a VP or an auditor. It closes the gaps this guide is about: the context nobody wrote down, the evidence nobody built, and the last mile to the work itself.

Live at 8 of the top 10 pharma companies · Five years a Gartner Magic Quadrant Visionary

---

BOOK A TAILORED DEMO

See an AI worker run on your workflow.

Scan the code, or visit tellius.com/demo. We will run this readiness assessment on one of your real workflows, with your people and your data reality.

tellius