

ADOPTION AND PERFORMANCE

Proof of Adoption

Evidenced AI adoption in operations centres: how to know AI is working, and prove it

Made by the team at Jaid

24 October 2026

Proof of Adoption

Executive summary

Most operations centres can tell you how many AI licences they bought. Very few can tell you how much of their work AI actually did last month, how well it did it, or what changed as a result. **That gap, between AI that is available and AI that is evidenced, is where most adoption programmes quietly fail.**

The research is blunt. MIT's 2025 study of enterprise generative AI found that the large majority of organisations saw no measurable return, and that the projects that worked were tied to a specific workflow, integrated into daily operations, and led by line managers rather than central AI teams. Meanwhile, ops workers adopt AI anyway, through personal tools that leave no trace on the case.

This paper sets out a different approach: **evidenced adoption**. It rests on three ideas:

1. **Measure adoption on the case, not the licence.** The unit of adoption is a client request handled with AI help, recorded with what the AI did, what the person changed and what the outcome was. Logins and prompts are not adoption.
2. **Climb a ladder, one workflow at a time.** AI earns more responsibility in stages, from triaging and silently testing, to assisting, to acting within limits. Each step up is a gate passed on evidence and signed off by the line manager and risk owner. Each step can be reversed automatically if the evidence weakens.
3. **Run product and client success as one programme.** The same evidence that tells a line manager their team is ready to move up tells us where to propose a better method, and which steps are running so reliably that they should be automated next. It also tells the client's board what they are getting for their money.

One Jaid client tripled its operations centre throughput without cutting headcount. Another cut missed client requests by 80%. Neither happened because AI was switched on. Both happened because adoption was planned, measured and earned.

1. Why AI adoption stalls in operations centres

Operations centres should be the easiest place to adopt AI. The work is high-volume, repetitive and text-heavy, and the pressure on capacity and SLAs never lets up. MIT's research even found that the strongest returns from generative AI come from back-office operations, not the sales and marketing tools that attract most budgets. Yet most programmes stall in the same five places.

1. **The pilot never meets real work.** A proof of concept runs on a curated sample, impresses a steering committee, and then meets the real mailbox: messy threads, odd attachments, request types nobody listed. MIT found that only a small fraction of custom enterprise AI tools ever reach production.
2. **The tool sits beside the work, not in it.** A general assistant in a side panel asks the ops worker to copy, paste, prompt and check. It helps with individual tasks, but the process itself does not change, so neither do throughput or SLAs.
3. **Nobody trusts what they cannot check.** An ops worker who cannot see where a figure in a draft came from will either redo the lookup or ignore the draft. A line manager who cannot see accuracy by request type will not let AI near anything that matters.

4. **The system does not learn.** MIT identified this as the core barrier: most systems do not retain feedback or improve over time. The same mistakes recur, and goodwill drains away.
5. **Success is measured in licences.** Programmes report seats provisioned and active users, because that is what the tools can count. The board asks what changed in the operation, and nobody can answer.

Underneath all five is the same gap: **adoption is being managed without evidence**. Nobody can show, request by request, what AI did, whether it was right, and what difference it made.

2. Adoption is not usage: what evidenced adoption means

Usage tells you that people opened a tool. Adoption tells you that the operation now runs differently. The two are easy to confuse, and most dashboards measure the first while claiming the second.

Usage metrics	Evidenced adoption metrics
Licences provisioned, active users	Share of client requests where AI did defined work
Prompts sent, sessions per week	Share of AI actions or AI handled cases approved without edits, by request type
Self-reported time saved	Measured handling time and SLA performance, before and after
User satisfaction surveys	Missed-work rate and client follow-ups
"AI was used on this team"	"AI did this on case 1042, using this evidence, and this person approved it"

We define **evidenced adoption** as AI use that is:

- **On the case.** Every AI contribution is recorded against the client request it served, not in someone's personal chat history.
- **Attributable.** You can see what the AI did, what the person changed, and who made the final decision.
- **Measured against outcomes.** Its value shows up in throughput, SLAs, missed work and quality, not just activity.
- **Earned by stage.** The responsibility AI holds in each workflow matches the evidence it has built up there.

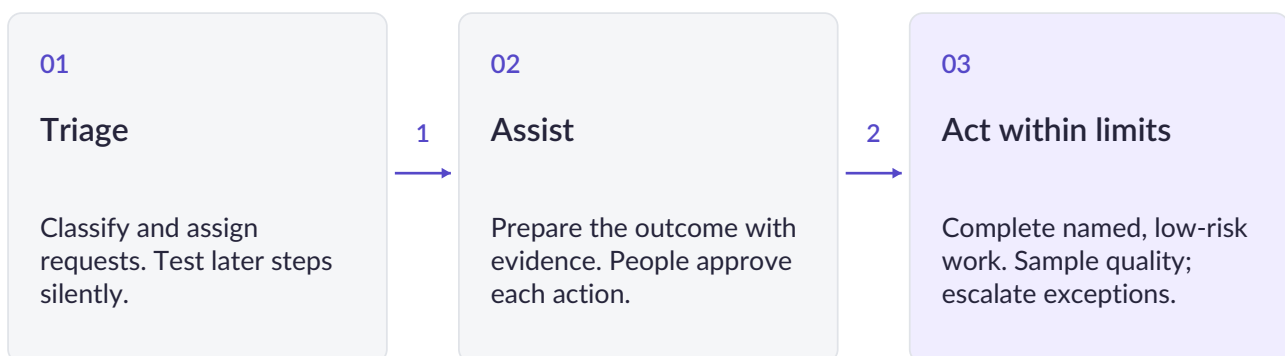
This definition matters because it changes who can say yes. A COO can commit to scaling AI that has evidence behind it. A risk owner can sign off a step that shows its own error rate. A line manager can give their team more AI help with confidence, because they can see where it works and where it does not.

3. The adoption ladder

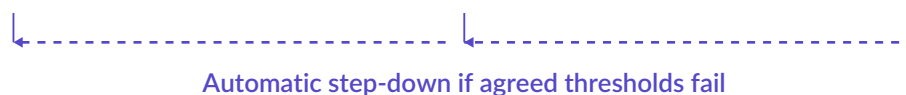
AI should earn responsibility the way a new team member does: by showing, on real work, that it can be trusted with more. We use a three-stage ladder. Each workflow, and often each request type within it, climbs independently.

1. **Triage.** From day one, AI triages each incoming request and assigns it to the right team or person, which helps the line manager straight away. Everything beyond that it tests silently: for each case it works out what it would do next and prepares the reply and the system actions, without anyone seeing or acting on them. Comparing those results with what ops workers actually did builds statistics on what the AI can already handle, request type by request type, and every mismatch becomes an example it learns from. By the time a workflow moves to Assist, the team has an honest baseline and the AI has already improved on their real work, with no risk to clients.
2. **Assist.** Ops workers see AI triage, extracted details and suggested next steps on each case, and AI prepares the complete outcome with its evidence attached: the reply to the client and the actions needed in other systems, such as updating a record, raising a task for another team or entering a transaction. People review and approve every one before it happens. The question is how often the prepared work goes through unchanged, and how often it is wrong.
3. **Act within limits.** For named, low-risk request types, AI completes the work, including the actions in other systems, once control checks pass, with sampled quality review rather than approval of every case. Anything outside the limits still goes to a person.

FIGURE 1 / The adoption ladder



Both gates require agreed evidence and sign-off by the line manager and risk owner.



Each workflow advances through evidence gates agreed by the line manager and risk owner; breached thresholds trigger a step down.

Responsibility grows only as fast as the evidence; the same thresholds that open a gate also trigger the step back down.

Three rules make the ladder safe:

- **Gates, not dates.** A workflow moves up only when it passes the evidence criteria for the next stage. A timetable is a plan; a gate is a decision.
- **Two signatures.** The line manager who owns the team and the risk owner for the workflow both sign off each step up. Both can see the evidence.
- **Automatic step-down.** If accuracy, override rates or complaints move past agreed thresholds, the workflow drops back a stage until the cause is fixed. Climbing is earned; falling is automatic.

4. The evidence model: what to measure at each stage

Each stage asks a different question, so each needs different evidence. The point is not to build the biggest dashboard. It is to collect exactly the evidence that the next gate will ask for.

Stage	The question	Evidence to collect	Typical gate to move up
Triage	Would the AI be right on our real work?	Triage and assignment accuracy against human decisions; silent-test results for replies and system actions, by request type	Accuracy agreed with the line manager and risk owner, held over several weeks of live traffic
Assist	How often is the prepared work right first time?	Approved-without-edit rate for replies and system actions; type and size of edits; errors caught at review; handling time and SLA performance	High approve-as-is rate for the request type; review catching few material errors
Act within limits	Does quality hold without case-by-case approval?	Sampled QA results; override and complaint rates; attempted actions blocked by controls	Sustained quality on samples; thresholds for automatic step-down agreed

The gate values are deliberately left to each client and workflow. A balance confirmation and a change to settlement instructions do not carry the same risk, and should not share a threshold. What must be consistent is that the threshold is **set in advance, written down and measured on the case record**.

Alongside the stage-specific evidence, four operational measures run throughout, because they are what the business actually cares about:

- **Throughput:** requests resolved per person per day.
- **Responsiveness:** time to first response and time to resolution against SLA.
- **Missed work:** requests not actioned or followed up in time.
- **Quality:** errors reaching clients, complaints and rework.

If the ladder is working, these move in the right direction as workflows climb. If they do not, the ladder is measuring the wrong things.

5. One playbook for product and client success

In most software companies, product builds features and client success drives usage, and the two meet at renewal time. For AI in operations that split does not work. Adoption depends on the product improving in exactly the places a client's work exposes, and the product only knows where those are if client success is reading the evidence with the client every week. So at Jaid they run as one programme, built on one shared evidence base: the case record.

The evidence loop

The case record does more than prove adoption after the fact. It is what tells us what to do next. Every week, the evidence from each workflow is read for three kinds of signal:

- **Steps that work almost perfectly.** When a request type has been approved without edits case after case, with no errors caught at review, it is a candidate to move up the ladder. We propose the step up, with the evidence, rather than waiting to be asked.
- **Steps that work, but could work better.** Where ops workers keep making the same edit, or the same lookup slows every case, there is usually a better method: a new data connection, a refined request type, a change to the review rule. We propose it, test it silently first, and let the evidence decide.
- **Steps that do not work yet.** Repeated overrides or misses on a request type point to a gap in the model's understanding. Those cases become training and evaluation examples in Jaid, so the next release handles them, and the case record shows whether it does.

FIGURE 2 / The evidence loop



Weekly reviews surface proposals. Monthly governance makes step-up decisions.

Weekly review proposes improvements; testing and a governance decision determine what changes.

Every improvement and every new piece of automation enters the same loop, and is judged by the evidence it produces.

This is how continuous improvement becomes concrete. Automation is never switched on because it seems like a good idea. It is proposed because the evidence shows a step is already running reliably, and it is approved through the same gates as everything else.

Who does what

Role	Responsibility in the loop
Client line manager	Owens the team's workflows, reviews evidence weekly, signs off each step up
Client ops workers	Do the work, approve or correct the AI, flag what it gets wrong
Client risk owner	Agrees thresholds in advance, co-signs each step up, receives step-down alerts
Jaid client success	Runs the weekly evidence review, proposes step-ups and better methods
Jaid product and AI teams	Turn recurring misses into training examples, ship improvements, re-run evals

The cadence

1. **Start with one workflow.** Scope it, connect the mailbox and systems, and agree the request types, thresholds and gate criteria before any AI runs.
2. **Triage and test silently for a few weeks.** Build the baseline on live traffic. Share the results openly, including where the AI is weak.
3. **Review weekly, decide monthly.** Weekly evidence reviews surface proposals; a monthly governance session with the line manager and risk owner makes the step-up decisions.
4. **Expand from what works.** Once one workflow is climbing, the same pattern is repeated for the next, reusing what was learned.

6. Line managers and ops workers: making the governed path the easy one

Adoption is decided on the floor, not in the steering committee. MIT's research concluded that line managers, not central AI teams, should lead AI integration into daily operations. Our experience agrees: the line manager who allocates work each morning, and the ops workers who handle it, are the people whose trust has to be earned.

For line managers, evidence replaces anxiety. Today a line manager is accountable for their team's output but cannot see what AI their people are using or how well it works. Evidenced adoption gives them a view they have never had: which request types AI handles well, where their team is overriding it, and where the backlog is building. They stop policing AI use and start directing it.

For ops workers, the governed path has to be the fastest path. People reach for personal AI tools because those tools are quicker than the official process. Bans do not change that; a better tool does. When the

triage is already done, the details are already extracted and the draft already carries its evidence, there is no reason to paste a client email into a chatbot.

Three practices make the difference:

- **Show the AI's working.** Every suggestion links to the evidence behind it. An ops worker who can check a figure in one click will trust the draft; one who cannot will redo the work.
- **Make corrections count.** When an ops worker fixes a draft, that correction feeds the evidence loop and, where it recurs, becomes a training example. People engage when they can see their feedback change the system.
- **Share the wins.** Weekly evidence reviews should report what the team achieved, not only what went wrong: requests cleared, SLAs met, missed work avoided. Adoption spreads when people see their colleagues getting time back.

The aim is not to replace the team. Our client that tripled throughput did it without cutting headcount. The same people now handle far more work, and spend their time on the cases that need judgement.

7. Case studies: what others learned

Two well-documented cases from financial services show both sides of the argument.

Morgan Stanley: evaluation first, then near-universal adoption. When Morgan Stanley Wealth Management rolled out its internal AI assistant to financial advisors, it began by building evaluations: structured tests of whether the AI's answers met the firm's standards, run before anything went live and extended as new uses were added. Only once the evals gave the firm confidence did the tool move into production. The result was adoption most enterprise software never sees: the firm reports that over 98% of advisor teams use the assistant. The lesson for operations centres is that **adoption followed evidence of quality**, not the other way round.

Klarna: scale without the right evidence. In 2024 Klarna announced that its AI assistant was handling two-thirds of customer service chats in its first month and doing the work of roughly 700 agents. A year later its CEO told Bloomberg that cost had been too dominant a factor in how the change was evaluated, that quality had suffered, and that the company was hiring human agents again. The AI was used at enormous scale; what was missing was evidence of quality, by type of request, deciding how far automation should go. The lesson is that **usage numbers can look like success right up until they don't**.

Read together, the two cases make the argument of this paper. Both firms deployed capable AI. One let evidence decide what the AI should do and earned the trust of its people. The other let volume and cost decide, and had to step back.

From Jaid's own clients. The same pattern shows up in operations centres. One client tripled its operations centre throughput without cutting headcount. A prime brokerage reduced missed client requests and follow-ups by 80%. An asset manager answers regulatory enquiries 97% faster. And Calastone's COO describes the value not only as capacity but as insight: a much clearer view of how the business is changing, how to plan capacity and how to meet client SLAs.

8. Questions to ask before your next AI rollout

Whether you are reviewing an existing programme or planning a new one, these questions show quickly whether adoption is being evidenced or assumed.

- ☐ **Can we say what share of last month's client requests AI actually worked on, by workflow and request type?**
- ☐ **For each workflow, do we know what stage AI is at, and what evidence would move it up or down?**
- ☐ **Were the thresholds for each step agreed in advance, and signed off by both the line manager and the risk owner?**
- ☐ **Is every AI contribution recorded on the case, with what the person changed and who approved it?**
- ☐ **Are we measuring throughput, SLAs, missed work and quality, or only logins and licences?**
- ☐ **When ops workers correct the AI, does anything change, or do the same mistakes come back next week?**
- ☐ **Does someone review the evidence weekly and propose what to automate next, based on what is already working?**

If most of the answers are "no", the programme is relying on hope. The good news is that every one of these can be put in place on a single workflow within weeks, and the evidence starts building from the first day.

Conclusion

AI adoption in operations centres does not fail because the technology is weak. It fails because firms switch AI on, count the licences and hope. The firms that succeed do something more disciplined: they let AI earn responsibility one workflow at a time, they record every contribution on the case, and they let the evidence decide what happens next.

That evidence does three jobs at once. It proves to the board that AI is changing the operation. It gives line managers and risk owners the confidence to say yes. And it shows exactly where to propose a better method, and which steps are already reliable enough to automate.

This is how we work with every Jaid client. We would welcome the chance to pick one of your workflows, put it on the ladder, and show you the evidence within weeks.

Sources

- [Computing: MIT report finds 95% of corporate generative AI pilots fail to deliver returns \(August 2025\)](#)
- [Virtualization Review: MIT report reveals the "GenAI divide" \(August 2025\)](#)
- [OpenAI: Morgan Stanley uses AI evals to shape the future of financial services](#)
- [InvestmentNews: Morgan Stanley's AI solution for advisors has expanded \(June 2024\)](#)
- [Fortune: Klarna is hiring humans again after AI push \(May 2025\)](#)
- [Mind the Product: Klarna goes on recruitment drive after ditching AI-first policy](#)

Contact: success@jaid.io