Triad

/EDITION 1

Triad • Reliability Report • Edition 1

The AI Agent Reliability Report for Property Management

/EDITION 1

The AI Agent Reliability Report for Property Management

Your team spends its days reading messages that need no answer, typing the same updates into the same systems, and juggling technicians' calendars. We build AI employees that take that work over: software that lives in your inbox, your ticket system and your planning, reads and decides like a trained colleague, and hands over to a person the moment it isn't sure. Before we put one in front of a property manager, we measured it. This report is that measurement, number by number, with how each number was counted.

BASIS OF MEASUREMENT

All results are test results, measured by Triad in its own test and validation environments, anonymised and aggregated; no figure in this report represents any specific client, customer, property or supplier.

/CONTENTS

In this edition

/01 An owner shouldn't have to take a vendor's word for it

/02 Nobody should be paid to read a message that needs no answer

/03 What stopped happening

/04 Nothing goes live that hasn't been right 19 times out of 20

/05 The AI doesn't get keys you didn't hand it

/06 A tenant shouldn't wait 11 minutes for what takes 2

/07 What you can hold us to

APPENDIX

/A How we counted

/B Where the outside numbers come from

/CHAPTER 01

An owner shouldn’t have to take a vendor’s word for it

AI can read a shared inbox, sort complaints and plan technicians. The harder question for any owner is simpler: can you rely on it? The doubt is real and measured. In AppFolio’s 2026 industry survey, 78% of property managers said they cannot yet rely on the AI features in their existing software. Yet the firms that did adopt AI expect their portfolios to grow 31% this year, nearly triple the growth expected by those that didn’t.[1]

That gap, between doubt and potential, is why we wrote this report. Most vendors will tell you their AI is accurate. We would rather show you the count: how many cases, over which period, checked by whom, and where it missed. Every figure below has a sample size and a confidence interval in the appendix. Where a number is a projection, it says so.

/CHAPTER 02

Nobody should be paid to read a message that needs no answer

Before the results, the problem in plain terms. In a property or facility operation, three kinds of manual work quietly eat the payroll:

/01 Reading. Most of what lands in a shared inbox needs no action at all, but someone still has to read every message to know that.

/02 Retyping. Every report becomes a ticket, a note, an update to the tenant, an update to the owner, a work order to a supplier. The same facts, typed four times.

/03 Puzzling. Which technician, which day, which route? Planning is a daily jigsaw solved by whoever has the least time for it.

None of this is skilled work. All of it is paid work. This is the work our AI employees remove, and here is what happened when they did.

/CHAPTER 03

What stopped happening

/3.1 Nobody reads the inbox anymore

In our property-management environment, we audited every single message the system handled over a four-day stretch, read back one by one by a person. The result: 95% needed no human answer at all. At that summer rate, a year of this inbox is 13,350 messages read, and only 623 that need a person. At one minute per message, that is roughly 210 hours a year of reading that no longer happens. The system read, sorted and routed the rest on its own: invoices, supplier threads, noise. And it judged correctly. In a separate window of 876 messages, it correctly ignored about 9 out of 10 junk messages while catching the one genuine emergency and flagging it as urgent. And because it is software, the inbox is read at 3 a.m. the same way it is read at 3 p.m.: no shift, no backlog on Monday morning.

FIGURE 1 · READING REMOVED

13,350
→ 623

messages a year, at the measured rate: of 13,350 read, only 623 need a human answer. The other 95% is reading work that simply stops existing. (Four-day census of 150 messages, projected to a year at that rate; the hours assume one minute per message.)

/3.2 Nobody plans the technicians anymore

In our field-service environment, the planning engine schedules the technicians: it knows every calendar, calculates travel times, and writes down, on every job, why it chose who it chose. Three moments from the planning log, one summer:

/01 The emergency. An urgent report came in through the emergency form. The engine had a technician on site the same morning, and moved a routine job aside to make room. Nobody touched a calendar.

/02 The 2½ hours nobody drove. Three technicians were available: 186, 56 and 34 minutes away. The engine sent the one 34 minutes away: 22 minutes closer than the next-best choice, 2½ hours closer than the worst. Multiply that choice by every job, every week: that is fuel, hours and mood you are not paying for anymore.

/03 The stolen keys. A tenant who couldn't lock her own front door. The engine picked the technician an hour closer and four days sooner than the alternative, then bundled a later job at the same address, so one trip resolved three tickets.

/3.3 Nobody types the same fact four times anymore

When a report comes in, the system sorts it into the right category (heating, leak, appliance), and that sorting decides who pays and who gets called. We tested this the hard way: the same cases, fed to the system again and again, across every category we handle. It gave the same, correct answer every single time: 444 runs across all 15 categories we handle, zero deviations. A person having a bad Tuesday does not do that. From that one decision, the ticket, the tenant reply in the tenant's own language, the owner update and the work order follow automatically.

/3.4 A support line shouldn't need a night shift

CASE: AN ENTIRE WHATSAPP SUPPORT LINE, RUN BY AI

In our customer-support environment, an entire WhatsApp support line runs on our AI employee: it answers day and night, sorts every complaint, opens the ticket, and when a customer deserves compensation, it issues the discount code itself. Compensation with a cost attached, handled by software, within limits it cannot cross.

Over one two-week stretch it handled about 215 conversations and tasks a day. Of roughly 3,000 runs, 3 stopped or stalled (99.9% run stability), and every one of those was caught by our own monitoring, not by a customer. Whether the answers themselves were right is measured separately, and that is where the money comes in: in twelve days, 36 customers got exactly one code each, nobody got two, and in 120 deliberately tricky cases, not one genuine customer was wrongly refused.

One more number a business owner will appreciate: when a better AI model became available, we switched, and cut the running cost of this system by a factor of twenty, from about 11 cents per conversation to well under one cent, measured over 4,939 conversations, with quality re-tested and held. Your automation gets cheaper as it matures, not more expensive.

/CHAPTER 04

Nothing goes live that hasn't been right 19 times out of 20

Every claim above survives because of how we work, not luck.
Three habits, in plain language:

/01 Nothing ships until it passes 19 out of 20. Before any change goes live, we run it against the same cases twenty times over. If it doesn't give the right answer at least 19 times out of 20, it does not ship. Full stop.

/02 Hundreds of automatic checks, on every change. When we adjust anything, the system re-tests itself against its own history. After one recent change we re-ran 588 past cases: not a single outcome changed that shouldn't have.

/03 A human audits the results. Periodically, a person separate from the build reads the outcomes back one by one, like the four-day inbox audit above, to confirm the system did what it should have.

Across the three control layers we measured, the weakest scored 99.38% and the pooled result over 1,638 runs was 99.69%: an error rate under 1% either way, with the full accounting in the appendix for anyone who wants to check our math. That is the point of this report: you don't have to take our word for anything in it.

/CHAPTER 05

The AI doesn't get keys you didn't hand it

The fastest way to lose money with AI is to give it keys it shouldn't have. So our systems come with hard limits, and one rule above all:

> You decide what it may send.
> Not us, and not the AI.

Whether the AI may send messages or create work orders on its own is a choice you make, per process, never a default. Out of the box, a person reviews everything before it goes out. Where you deliberately choose autonomy (say, routine tenant confirmations), the sending is done by fixed, predictable software following an approved plan, with an automatic content check, never by the AI free-styling.

/01	An unknown supplier never gets a work order; you get a notification instead.
/02	Costs above the mandate you set trigger an approval request first.
/03	When the system isn't sure, it stops and hands the case to a person with a summary; it does not guess.

These aren't promises; they're measured behaviours. The compensation guardrails in the support-line case above are those limits working, with money on the line.

/CHAPTER 06

A tenant shouldn't wait 11 minutes for what takes 2

Independent research across customer service broadly shows what operators feel daily: published industry figures report AI resolving requests in 1.9 minutes on average against 11.4 minutes for human handling, with error-complaint rates around a third of a percent.[2] Different processes than ours, not directly comparable, but the direction is unmistakable. The businesses that move first get the compounding benefit: every month of automated intake, sorting and planning is a month of payroll spent on work that matters instead.

> The work your team shouldn't be doing is already gone at ours.

/CHAPTER 07

What you can hold us to

If you work with Triad, these are not aspirations. They are the rules the numbers above were produced under, and they apply to your implementation on day one:

/01 Nothing goes live below 19 out of 20. Every process is run against your own cases, twenty times over, before it touches your inbox, your tenants or your suppliers.

/02 Every number comes with its count. You get the same accounting you just read: sample size, window, misses. Monthly, for your own operation.

/03 A person audits the output. Someone separate from the build reads real outcomes back, on a schedule, and the misses are reported to you, not buried.

/04 You decide what it may send. Autonomy is switched on per process, by you, never by default. Unknown suppliers, costs above your mandate and low-confidence cases always go to a person.

This report is the first in a series; every next edition adds new implementations and fresh numbers. If you want to know what an AI employee would take off your team's plate, and hold us to the same standard of proof you've just read, the conversation is free.

Talk to Triad ↗

triadagency.ai/contact

Triad

/APPENDIX A

How we counted

Every figure in this report, with its sample size, measurement window and 95% confidence interval (Wilson method). A figure without an interval is an estimate posing as a fact. Sample sizes below are the actual measured counts; the annual volumes in the body are these measured rates projected to a year, and are labelled as projections wherever they appear.

TABLE 1 · EVERY FIGURE, WITH ITS COUNT

FIGURE (AS USED IN THE BODY)	N	RESULT	95% CI	HOW MEASURED
Inbox census: share needing no human answer	150	95.3%	90.7 — 97.7%	every message over 4 days, read back by a person separate from the build; 143 of 150 needed no action, 7 did
Inbox census: routed correctly	150	99.33%	96.32 — 99.88%	same census; 1 genuine miss, rule fixed same week
Junk filtering	876	≈ 90%	87.79 — 91.77%	separate window; junk messages correctly ignored, 1 genuine emergency in the window, flagged as urgent
Classification consistency	444	100%	99.14 — 100%	the same cases repeated across all 15 categories, tested before rollout; 0 deviations
Re-run after a change (regression)	588	100%	99.35 — 100%	past cases re-run after one change; 0 outcomes changed that should not have
Reporter recognition (intake)	59	88.14%	77.48 — 94.13%	gross; 3 of 7 misses were external parties correctly left unmatched (net 93%)
Intake: agent decision layer	320	99.38%	97.75 — 99.83%	measured runs; both errors caught by automatic retry
Support line: run stability	≈ 3,015	99.90%	99.71 — 99.97%	consecutive runs over 2 weeks, 3 incidents (runs stopped or stalled), all self-detected; measures completion, not answer quality
Support line: end-to-end process	697	99.57%	98.74 — 99.85%	measured runs, 2-week window, 3 errors
Support line: customer lookup step	621	100.00%	99.39 — 100%	measured runs, 2-week window, 0 errors
Support line: compensation guardrails	120	0 wrongly refused	n/a	two independent 60-run adversarial rounds; 36 codes issued in 12 days, 0 duplicates
Support line: cost per conversation	4,939	€0.11 → under €0.01	n/a	running cost before and after the model switch, quality re-tested on the same cases
Pooled control layers	1,638	99.69%	99.29 — 99.87%	end-to-end process + agent decision layer + lookup step; worst-case error 0.71%, under the 1% stated in the body

/APPENDIX B

Where the outside numbers come from

/01 AppFolio, *Property Manager Benchmark Survey 2026*, appfolio.com/newsroom/property-manager-benchmark-survey-2026. Figures verified against the live page, 26-08-2026. The 78% concerns AI features in legacy property-management software.

/02 Digital Applied, *Customer Service AI Agent Statistics 2026* (aggregation of Zendesk, Salesforce, Gartner, Forrester, Intercom and McKinsey research), digitalapplied.com/blog/customer-service-ai-agent-statistics-2026-data. Verified 26-08-2026.

/03 On measurement method: *Towards a Science of AI Agent Reliability*, arXiv:2602.16666, ICML 2026, the research position this report's repeated-measurement approach follows. Verified 26-08-2026.

NOTE ON MEASUREMENT

Sample sizes and measurement windows are stated per figure in Appendix A; annual volumes in the body are measured rates projected to a year. All results are test results, measured by Triad in its own test and validation environments, anonymised and aggregated; no figure in this report represents any specific client, customer, property or supplier.