ZTS ADVISORY

FIELD STUDY // 55 MYSTERY-SHOP CALLS // JULY 2026 // REVISED & BENCHMARKED EDITION

The Front Door Integration Test

Where PE-backed home services platforms prove, or fail, the thesis they paid for. Fifty-five calls, one rubric, four evidence labels, and every denominator in view.

0/55

live sibling referrals, in 13 calls where a second problem was volunteered unprompted

1/55

written recaps received before hang-up; 36 of 49 answered calls never asked for an email

7.6×

the rate at which scripted agents attempted the save versus humans ($p = 0.004$)

THE QUESTION, AND THE RULES OF EVIDENCE

Does the front door pass the test the deal model already assumes it passes?

In July 2026 I placed 55 mystery-shop calls to 48 residential service brands owned by roughly 20 private-equity sponsors. Every call was transcribed and scored against one eight-stage rubric. This report is the answer, rebuilt so it survives your data team: externally benchmarked, statistically bounded, and labeled line by line. All brands and sponsors are anonymized; verbatim call quotes carry the argument throughout, cleaned only of transcription artifacts.

EXHIBIT 1 · THE EVIDENCE LADDER

Every major number in this report carries one of four labels.

LABEL	WHAT IT MEANS	EXAMPLE
OBSERVED	Directly found in the 55-call field study. Comes with a numerator and a denominator, always.	0/55 live sibling referrals
EXTERNAL BENCHMARK	Found in a cited external dataset or authoritative publication. The bibliography states each source's sample, method, and definitions, and flags vendor research.	Home services converts 46% of phone leads during the call (Invoca, 60M+ calls)
INTERPRETATION	My reasoned explanation of the evidence, with the competing interpretation stated next to it.	"Config compounds; coaching decays"
MODELLED	Calculated using disclosed assumptions; every input is observed, benchmarked, or flagged for replacement.	Base-case $409K EBITDA opportunity

Contents

HOW TO READ THIS REPORT

Each finding runs two pages and reads standalone: the exact claim, the evidence label, the denominator, one verbatim quote worth reading aloud, the external benchmark, the alternative explanation, and a box you can act on Monday, now including a controlled pilot, a KPI, a guardrail, and a decision rule. Forward a single finding to the operator it belongs to; the page furniture tells them where it came from.

What the study can and cannot prove. It observed intake behaviors, never outcomes: no completed bookings, completed jobs, competitor wins, revenue, or EBITDA, and no claim that a missed behavior caused a lost job. What it can prove is narrower and more useful: whether the behaviors that selling standards, platform documentation, and conversion research all treat as necessary were present on the calls the marketing budget paid to generate.

SUMMARY

The two-page version

I placed 55 mystery-shop calls to 48 residential HVAC, plumbing, and roofing brands owned by roughly 20 private-equity sponsors, between July 16 and 25, 2026. I presented as the caller every one of these businesses says it wants: a homeowner with a real problem, money to spend, and two or three companies on the list. Every call was transcribed and scored against one eight-stage rubric.

The central thesis. In PE-backed home services, the front door is the real integration test: paid demand turns into booked revenue only when every brand uses consistent rules to recognize intent, secure a next step, route value, and measure the outcome. Administrative integration (shared ERP, shared price book, shared board pack) shipped at most of these platforms. Customer-facing integration mostly did not, and the caller is the only person positioned to notice.

The five strongest observed findings

1. In 13 calls I volunteered a second problem in the home, unprompted. Across all 55 calls, no human or AI agent made a live referral to a sibling brand, in any phrasing. OBSERVED 0/55; 95% CI upper bound 6.5%.
2. Four sibling brands under one sponsor, shopped inside 48 hours with the same scenario, delivered four materially different front doors. The diagnostic fee was identical everywhere; recap, capture, second-problem handling, and membership behavior were local habit. OBSERVED n=4 brands.
3. Scripted AI agents attempted to save a deferring caller in 6 of 13 agent-run calls; human answerers attempted it in 2 of 33. The difference is statistically significant despite the small sample (Fisher's exact p = 0.004). OBSERVED
4. Five calls presented a competitor's written $8,500 replacement quote and offered to email it. Zero platforms asked to see it, and soonest visits ran two to ten days for a buyer who announced a weekend decision. OBSERVED 0/5; small denominator, wide uncertainty, treated accordingly.
5. Sixteen calls were fronted by AI agents. The failure modes observed (fake email accepted after being announced as fake, a 150-mile address misvalidation, an unescalated empty-calendar loop) are configuration failures, and configuration executes identically on every call. OBSERVED within n=16.

Three externally supported implications

1. **Cross-sell is the hardest synergy class to land.** McKinsey's survey of 200 M&A executives found most companies undershoot revenue-synergy targets, by an average of 23 percent, and separate McKinsey research attributes roughly 20 percent of revenue-synergy value to cross-selling specifically. A cross-sell thesis with no front-door machinery is priced-in value with no delivery mechanism.
2. **Speed decides high-intent leads.** The Oldroyd/InsideSales research and the 2011 HBR audit of 2,241 firms established that response inside an hour multiplies qualification odds roughly sevenfold versus waiting longer. A two-to-ten-day first visit for a quote-in-hand replacement buyer sits on the wrong side of every curve in that literature.
3. **The measurement gap closes inside owned systems.** ServiceTitan's Contact Center Pro documentation describes call reasons, agent performance reporting, abandoned-call reporting, and voice-agent escalation configuration; Genesys Cloud documentation defines service level, abandonment, and short-abandon exclusions. Nothing in this report's fix list requires software the platforms do not already run, though several capabilities depend on product edition and configuration.

Three actions

1. Run the Monday-morning self-test this week. Seven questions cost one hour of listening to recorded calls; three are data pulls. Formulas are published in full, nothing redacted.
2. Stand up one controlled pilot in the next 30 days. The written-recap randomization is the cheapest and cleanest.
3. Put call-to-booked-job, on one definition across brands, into the board pack next to marketing spend. The dashboard section gives the numerator, denominator, exclusions, and owner for every metric.

WHAT THIS STUDY IS, AND IS NOT

It can prove that specific revenue-relevant behaviors were absent or inconsistent across a structured sample of 48 brands, with exact counts and confidence intervals. It cannot prove conversion impact, because no bookings were completed; the economic section is therefore a scenario model with disclosed assumptions, never a measurement. The sample is a structured control-failure detection study, and I do not claim it is statistically representative of the industry. What it detects, your own call recordings can confirm or refute inside a day.

EXHIBIT 2 · THE CLIFF, OBSERVED

From forty-nine answered calls to zero referrals, in four steps.

49/55

answered by a live voice. The phone system did its job.

8/49

tried to hold the lead when I deferred. Six were machines.

1/55

put anything in writing before I hung up.

0/55

referred a sibling brand, live on the phone, including all 13 calls where I volunteered a second problem unprompted.

Observed, July 2026. The first number runs on software the platform bought. The next three are workflow. Denominators vary because six calls never reached a voice and some stages trigger only when the call creates their condition.

CONTEXT

Why the front door matters to the investment thesis

The consolidation math is public and it is aggressive. Kroll's residential HVAC coverage estimates the global HVAC market above $350 billion in 2025 with mid-single-digit growth ahead, and judges the residential consolidation cycle only about halfway run. Capstone Partners tracked private-equity add-on activity in HVAC services rising 88 percent year over year through mid-2025, with financial buyers taking roughly half of all transactions. Multiples at the platform tier assume operational transformation, and the deal decks assume specific mechanisms: shared demand across trades, cross-selling between sibling brands, one customer base served by many flags, and digital operating leverage from a common software stack. EXTERNAL BENCHMARK

Here is the problem the deal decks skate past. Revenue synergies are the synergy class that most reliably disappoints. McKinsey surveyed 200 executives who had each led revenue-synergy programs on multiple deals above $2 billion; the majority reported falling short of their revenue-synergy aspiration, with an average gap of 23 percent between goal and attainment, and revenue synergies took materially longer to capture than cost synergies. Separate McKinsey work found cross-selling accounts for roughly 20 percent of the value companies derive from revenue synergies, and that companies failing to pursue revenue synergies alongside cost synergies see combined-company sales growth fall by an average of seven percentage points against industry. EXTERNAL BENCHMARK

So a multi-trade roll-up thesis is, in large part, a bet on the hardest synergy class, and the bet has a physical location. The place where a homeowner would first experience "one platform, many trades" is the phone. Invoca's 2025 benchmark analysis of more than 60 million calls found home services carrying the highest phone conversion rate of any industry it measures, at 46 percent of phone leads converting during the call, against a 37 percent cross-industry average. That is vendor research with a disclosed sample, and its definitions differ from booked jobs; the definitional note before Finding 01 handles the difference. The channel where the industry converts best is also the channel this study found least governed. EXTERNAL BENCHMARK

One distinction organizes everything that follows. Administrative integration is what the integration memo tracks: ERP migration, payroll, insurance, the price book, the board pack. Customer-facing integration is what a caller experiences: whether intent gets recognized, whether a next step gets secured, whether value gets routed to the right person, and whether anyone measures the result. In this sample the first kind had visibly shipped, and the second kind mostly had not. My claim, labeled INTERPRETATION and defended throughout: the gap between those two integrations is where a measurable share of the priced-in thesis currently sits unbuilt.

METHOD

Method and limitations, limitations first

You diligence companies for a living, so the weaknesses of this study go first.

Design. 55 mystery-shop calls to the published consumer lines of 48 residential service brands (HVAC, plumbing, roofing) under roughly 20 private-equity sponsors, in the United States and Canada, July 16 to 25, 2026, across dayparts including evenings and a weekend. A small cast of homeowner personas used shopper-controlled phone numbers and inboxes. Four scenario families: failing hot water in an aging tank; an underperforming or dead AC, usually with a second problem volunteered mid-call; a ceiling stain on a 15-to-20-year-old roof; and a competitor's $8,500 full-system replacement quote, with an offer to email the proposal. Personas presented as comparison shoppers who defer at the close. That deferral is deliberate: without it, the hold, capture, recap, and follow-up stages never trigger, and those stages were the point of the study. Every call was transcribed; scoring was done from transcripts against the eight-stage rubric; quotes were cleaned only of transcription artifacts. Some brands received a second or third call where a first attempt dropped, a promised callback arrived, or an after-hours line warranted a separate test.

What this design is. A structured control-failure detection study. It asks whether specific, definable behaviors occur when a caller creates the exact conditions those behaviors exist for. It is the audit logic of a mystery shop, pointed at an operating layer.

What this design is not. A representative sample of industry conversion. I did not draw brands randomly from a defined population, per-brand reads rest on one to three calls, and no outcome past the phone call was observed. One call proves nothing about one brand. Fifty-five calls with the same misses repeating across 48 brands and roughly 20 sponsors is a pattern worth investigating in your own data, and that is the strongest representativeness claim this report makes.

Known limitations, enumerated. The shopper never books, so nothing here is a conversion claim. Eleven calls ended with a promised callback; two arrived, both same-day, both observed; the rest fell outside the study window, so no callback-kept rate is published. July is peak season, several brands were visibly slammed by a heat wave, and capacity excuses are real; note what capacity does not explain, since asking for an email address costs no truck. Scenario design inflates some incidences by construction (second problems were volunteered deliberately, replacement-quote calls were a designed test), and the economic model corrects for this explicitly. Per-brand reads are directional only.

Ethics. No bookings were confirmed, no technician was dispatched, no free work was extracted beyond prices and process information any caller may request. Where a rep offered to hold a slot, I declined or released it on the call. Two scheduled callbacks were accepted and taken. Brands and sponsors are anonymized here and in any excerpt.

EXHIBIT 3 · STATISTICAL UNCERTAINTY, IN PLAIN ENGLISH

Small samples carry wide intervals, and I publish them rather than hide them.

Exact (Clopper-Pearson) binomial 95% confidence intervals on the aggregate counts:

OBSERVATION	RATE	95% CI
Answered by a live voice, 49/55	89%	78% to 96%
Stated a visit price, 40/49	82%	68% to 91%
Save attempt at deferral, 8/49	16%	7% to 30%
Asked for an email, 13/49	27%	15% to 41%
Written recap before hang-up, 1/55	1.8%	0.0% to 9.7%
Live sibling referral, 0/55	0%	0.0% to 6.5%
Asked to see competitor's proposal, 0/5	0%	0.0% to 52%

Read the two zero rows differently, because they deserve it. Zero of 55 is informative even at zero events: the exact upper bound says that if the true referral rate across comparable calls were even 7 percent, seeing none in 55 tries would be unlikely. Zero of 5 is suggestive only; with five trials, a true ask rate as high as half the time is still consistent with observing zero. Finding 04 is therefore stated as a hypothesis your own tagged-call data can test in an afternoon, and never as a proven industry rate.

For the agent-versus-human save comparison, Fisher's exact test on 6/13 versus 2/33 gives p = 0.004 (odds ratio 13.3), so that contrast survives formal testing despite the sample size; the confidence intervals on each side remain wide (19% to 75% for agents, 1% to 20% for humans). Treat the 7.6× point estimate as a direction with magnitude uncertainty, and the existence of the gap as well supported.

Not yet collected or not yet published: call-level distribution by trade and daypart (held in the working dataset); a second-coder review with agreement statistics; regional identifiers (withheld for anonymity); membership-eligibility ground truth per brand. A published supplement should add the exact scripts, the distribution tables, the scoring codebook, call-level de-identified data, selected transcripts, and second-coder agreement. Until it does, treat single-coder scoring as a limitation. The companion supplement, including the full evidence matrix and replication design, is available on request at zaha@ztsadvisory.com.

THE FRAMEWORK

The Front Door Integration Test

The study scored calls on an eight-stage rubric. The rubric is the diagnostic instrument, and it stays. But operating partners do not manage eight stages; they manage controls. So the executive frame is five controls, and every stage, finding, and fix in this report maps to exactly one of them.

REACH

Can the caller reach someone who can price, schedule, or commit to a next step? Answering the phone is necessary and insufficient; eleven calls in this sample reached a voice that could neither price nor book. Reach failures live in routing tables, staffing models, and overflow rules.

RECOGNIZE

Does intake recognize what just walked in the door: urgency, replacement intent, membership eligibility, a second service need? Recognition is the difference between a $99 ticket process and an $8,500 decision process. In this sample, recognition machinery was the emptiest layer of all.

RETAIN

Does the platform secure a next step before the call ends: a booking attempt, a save attempt at deferral, contact capture, a written recap, a scheduled follow-up with a time on it? Retention is where 8/49, 13/49, and 1/55 live.

ROUTE

Does the caller reach the correct estimator, trade, sibling brand, or escalation path? Routing is where 0/55 lives, and where the AI agents' unescalated dead loops live.

REVIEW

Does management measure call-to-booked-job, booking quality, follow-up completion, and performance on one definition across brands? Review is the control that makes the other four durable. No platform in this sample could have produced its AI agent's call-to-booked-job rate on the human definition.

EXHIBIT 4 · EIGHT STAGES UNDER FIVE CONTROLS, OBSERVED

The software you bought works. The selling stops at stage four.

The stages are conditional: a stage only triggers when the call creates its condition, which is why denominators vary and why a simple sequential funnel misleads.

STAGE	OBSERVED	BAR	CONTROL	NOTE
1 · Answer	49/55		REACH	Six dead ends, all at brands with live paid media that week.
2 · Qualify	.		RECOGNIZE	The strongest stage. One weekend emergency still left without a phone number.
3 · Price	40/49		REACH	Stated the visit fee, or free. 9 could not price a visit; 11 calls hit answer-only staffing.
PAID SOFTWARE ENDS HERE · WORKFLOW BEGINS				
4 · Hold	8/49		RETAIN	Tried to keep me at the moment I deferred. Six of the eight were scripted AI agents.
5 · Capture	13/49		RETAIN	Asked for an email. 8 real; 1 announced out loud as fake, and accepted.
6 · Recap	1/55		RETAIN	A written recap received before hang-up. The best call in the sample, and the only one.
7 · Follow-up	2 obs.		RETAIN	Two promised callbacks observably arrived, of eleven promised. The only follow-up scheduled to a stated time was booked by a machine.
8 · Route	0/55		ROUTE	Live sibling referrals. 3 of 13 volunteered second problems got one-trip handling.

The shape of the table is the report's structural finding: stages that run on already-purchased software mostly work; stages that require configured workflow and management review are close to absent. Read down the Control column and the pattern is starker still: REACH mostly passes, RETAIN and ROUTE mostly fail, and REVIEW is unmeasurable from the outside because, on the evidence of what platforms could not answer, it barely exists on the inside. OBSERVED INTERPRETATION

A NOTE ON DEFINITIONS BEFORE THE FINDINGS

External benchmarks use terms that sound interchangeable and are not. **Lead conversion** (Invoca) means a phone lead showing a conversion signal during the call, measured by AI signal detection on Invoca's customer base, a self-selected population sophisticated enough to buy call analytics. **Call booking rate** (ServiceTitan) means qualified lead calls converted to booked jobs as classified by call reasons; ServiceTitan's own 2022 data put the average trades shop at 42 percent, and 59 percent for shops with 25 or more technicians. **A booked job** is an appointment that can still cancel. **A completed job** is revenue. When this report compares an observed behavior to an external rate, it names which definition is in play. Invoca's finding that only 35 percent of agents ask leads to purchase or book is the closest external cousin to this study's Hold stage, and even that pairing is imperfect: Invoca measures any ask on any lead call, while my Hold stage measures a save attempt at the specific moment a caller defers, a narrower and harder behavior.

01

The cross-sell thesis lacks front-door machinery

The exact claim OBSERVED : In 55 calls, including 13 in which the shopper volunteered a second problem in the home unprompted, no human or AI agent made a live, front-line referral to a sibling brand in any phrasing. Of the 13 volunteered second problems, 3 received one-trip handling by the same brand; the remainder drew a second visit, a second fee, or a redirection to a competitor.

0/55

calls with a live, front-line referral to a sibling brand, in any phrasing, by any human or agent. 95% CI upper bound: 6.5%.

13

calls where I volunteered a second problem in the home, unprompted. A shower losing pressure. Gutters. A water heater beside a dying AC.

3 of 13

second problems that got one-trip handling. The rest: a second visit, a second fee, or "call someone else."

> "We only handle hot water tanks or tankless water heaters, but I don't have any plumbers on staff for anything else. So you'd have to call a plumber for that, unfortunately."

CSR AT AN HVAC BRAND WHOSE PLATFORM SELLS PLUMBING UNDER OTHER FLAGS

External support EXTERNAL BENCHMARK : McKinsey's revenue-synergy research is the uncomfortable frame here. Cross-selling contributes roughly 20 percent of realized revenue-synergy value; most acquirers undershoot revenue-synergy targets by an average of 23 percent; and the named stumbling blocks in that research read like a description of this sample: front-line staff lacking the knowledge, capacity, and incentives to sell the wider portfolio. The external literature does not prove my calls are representative. It proves that when cross-sell fails, it fails exactly here, at the front line, for exactly these reasons.

Two scope notes, so this claim survives a pull of the logs. First, sibling brands often sit in different metros, so the finding concerns the mechanism, and never the miles: across 55 calls I never heard a transfer number, a warm handoff, a "my colleagues at our sister company," or a capture of the second need for later work. Second, this measures the live phone layer; if a platform routes cross-brand work through backend dispatch after the call, that machinery never surfaced to a caller in this sample, and a caller told "call a plumber" does not wait around to discover it.

The alternative explanation, stated fairly. Some platforms may deliberately keep brands commercially separate for licensing, service-area, or brand-equity reasons, in which case zero referrals is policy, and the finding becomes a different question: whether the deck's cross-sell paragraph should survive diligence at all. A second alternative: cross-sell may run through membership renewals and technician upsell in the home rather than through intake. Plausible, unobserved by this study, and testable in your own attach data.

Management implication INTERPRETATION : A synergy that cannot be dialed and cannot be measured does not exist operationally, whatever the deck says. The absence found here is an absence of machinery (routing rule, script line, lead-source code per sibling), and machinery is buildable.

CHECK THIS AT YOUR OWN PLATFORM MONDAY

- Pull one week of recordings for two sibling brands in one metro. Count calls where a second need is voiced; count live referrals to the sibling.
- Check whether a lead passed between siblings can even be tracked: a transfer path that exists, a lead-source code per sibling. If it cannot be dialed and cannot be measured, it does not exist.
- Ask each brand president for the referral script line. Time the silence.

CONTROLLED PILOT · ROUTE

In one metro pair, arm intake with three referral treatments across alternating weeks: warm transfer, scheduled sibling callback, and referral number with a lead-source code. Track referral attempts, completed referrals, and booked sibling jobs against a no-treatment sibling pair.

KPI: Sibling-referral completion rate on eligible calls (a second need voiced, or an out-of-trade need identified). **Guardrails:** Primary-job booking rate on referral-eligible calls, so referral pushing never costs the first job; caller handle time. **Decision rule:** Scale the winning treatment if referred-and-booked sibling jobs exceed 5 percent of eligible calls over eight weeks with a flat guardrail; kill treatments that move the guardrail more than two points.

02

Same platform, four different front doors

The exact claim OBSERVED : Four sibling brands under one sponsor, all consumer HVAC, shopped inside 48 hours with the same scenario, delivered identical diagnostic fees ($169, with a member discount offered) and divergent everything else.

EXHIBIT 5 · FOUR SIBLING BRANDS, ONE PLATFORM, 48 HOURS

	BRAND R1	BRAND R2	BRAND R3 · BEST DOOR	BRAND R4 · WORST DOOR
Diagnostic fee	$169, member discount offered	Same	Same	Same
Soonest visit	Tomorrow morning	Friday	Same evening	"Nothing after 4 p.m. for the foreseeable future"
Second problem	Bookable, second trade, second fee	"You'd have to call a plumber"	Bookable, separate appointment and fee	Declined, no referral
Membership	Scripted pitch with terms and savings math	Explained on request	Offered with terms	Never mentioned
Send it in writing?	No; waited while I found a pen	"No way to send you the information"	Texted a recap during the call	"Check the website"
Contact captured	Name, address, phone	Address only	Phone	Nothing

Observed, one platform, same scenario, inside 48 hours. Fee standardization held; everything after the fee was local habit. One call per brand: an existence proof that intra-platform spread can be as wide as industry spread, never an estimate of average spread.

> "I don't have any way to send you the information, but you can check it out on our website."

CSR AT BRAND R4, SAME PLATFORM, SAME WEEK THAT SIBLING BRAND R3 WAS TEXTING RECAPS MID-CALL

External support EXTERNAL BENCHMARK : The pattern (pricing standardized, behavior local) matches what the integration literature predicts: cost-side and administrative levers land early and reliably, while behavior-dependent revenue levers lag by years. The fee card is a configuration artifact; the recap is a habit. Config shipped. Habit did not.

The alternative explanation. One of the four calls may have caught an untrained new hire, a bad day, or a genuinely slammed shop; this was a heat-wave week. True, and partially beside the point: a system that lets a bad day erase capture, recap, and referral entirely is exactly the finding. The control question is variance, and variance is what an operating layer exists to remove.

Management implication INTERPRETATION : For a new CRO or COO this is the highest-yield diligence in the building, and the standard board pack hides it, because call-to-booked-job is either unreported or reported per brand on whatever definition each brand inherited. Adoption of the intake standard gets assumed in the integration memo at week twelve and never re-measured. The spread inside one platform, on identical software and an identical fee card, measures how much standardization actually shipped.

CHECK THIS AT YOUR OWN PLATFORM MONDAY

- Pull call-to-booked-job by brand for the trailing 90 days, on one definition, from the ServiceTitan Calls report. If the definitions differ by brand, that finding outranks the number.
- Have one person call each of your brands with the same two-problem scenario, same week, and score the calls against Exhibit 4. The delta between your best and worst front door is your integration backlog, ranked.
- Ask for the intake one-pager each brand's CSRs actually use. Count the versions.

CONTROLLED PILOT · REVIEW

Take your worst-scoring front door and your best. Transplant the best door's intake configuration and scripts to the worst for 60 days, changing nothing else. This is the closest thing to a controlled test of standardization value a platform can run without new spend.

KPI: Eight-stage rubric score by brand, monthly, from a fixed-size QA sample. **Guardrails:** CSR attrition and handle time at the transplant site. **Decision rule:** If the transplant site's rubric score converges to within 10 points of the donor site inside 60 days, standardize platform-wide on the donor configuration; if it does not, the blocker is management adoption rather than configuration, and the fix escalates from ops to the brand president's scorecard.

03

Scripted agents attempt the save more consistently

The exact claim OBSERVED : At the moment the shopper deferred, scripted AI agents made a live save attempt (hold the slot with free cancel, book as comparison backup, priority list, or a callback scheduled to the minute) in 6 of 13 agent-run calls. Human answerers made a comparable attempt in 2 of 33.

EXHIBIT 6 · WHO TRIES TO SAVE THE LEAD, WITH UNCERTAINTY

	OBSERVED	RATE	95% CI	BAR
Scripted AI agents	6 of 13	46%	19% to 75%	
Human answerers	2 of 33	6%	1% to 20%	

A save attempt: holding a slot, booking a backup, a priority list, or a callback with a time on it, offered at the moment the shopper deferred. Rate ratio 7.6×; Fisher's exact p = 0.004; odds ratio 13.3. The gap is statistically significant; the exact multiple carries wide uncertainty; the direction is well supported. The prior edition of this study titled this finding "the scripted agents out-close the humans." No closes were observed. The claim is about attempts, and it stays about attempts.

"Would it help to just go ahead and get us on the schedule as a free backup comparison?"

AFTER-HOURS AI AGENT AT A ROOFING BRAND, IMMEDIATELY AFTER I SAID I NEEDED TO TALK TO MY HUSBAND

The two human saves deserve naming, because they were the best selling in the sample: one rep held that evening's slot with text-to-cancel and no commitment; another waived the visit fee unprompted and offered to book so I "wouldn't have to do the process all over again." Both moves are one sentence long. Neither appeared anywhere else in 55 calls.

External support EXTERNAL BENCHMARK : Invoca's 60-million-call analysis found only 35 percent of agents ask leads to make a purchase or book an appointment, on a broader definition than my save-attempt measure. Both datasets, one enormous and one small, point the same direction: the ask is the industry's rarest cheap behavior. ServiceTitan's Contact Center Pro documentation confirms the mechanism is configurable, describing AI virtual agents that book jobs, handle overflow and after-hours calls, and escalate to live agents per configured rules; capabilities depend on edition and configuration.

The alternative explanation. Agent-fronted calls skew after-hours, and after-hours callers may present differently; the comparison is between call populations that differ in more than the answerer. Also, six of eight saves coming from agents partially reflects where platforms chose to deploy agents. Both caveats are real. Neither explains why the daytime human team, which takes the overwhelming majority of volume, almost never says the twenty words the platform already wrote down for its machine.

Management implication INTERPRETATION : Config compounds; coaching decays. The agent's save line was written once, by whoever configured it, and now executes at 9 p.m. on every call. The human save depends on which of forty CSRs picked up and what their last coaching session covered. The platform's best sales behavior already exists, in a configuration file it owns; it has simply never been deployed where the volume is.

CHECK THIS AT YOUR OWN PLATFORM MONDAY

- Pull ten after-hours and ten daytime recordings from your CCaaS. Count save attempts per ten calls, agent versus human. You now own the same exhibit I do, built on your own data.
- Find the save line in your AI agent's configuration. Ask when a daytime CSR last said those words on a live call.
- Ask who reviews the agent's unbooked calls, and on what cadence. "Nobody, never" is the common answer, and the cheapest fix in this report.

CONTROLLED PILOT · RETAIN

Train two matched daytime CSR cohorts; give one the agent's save script with a one-morning huddle and a wallet card, leave the other on current practice. Compare configuration periods if cohorting is impractical. Run 30 days.

KPI: Save-attempt rate on deferring calls, from QA sampling. **Guardrails:** Complaint rate and booking-cancellation rate, because a pushy save that books junk is worse than no save. **Decision rule:** Scale the script platform-wide if the trained cohort's save-attempt rate exceeds control by 15 points with flat guardrails; treat any rise in cancellations above two points as a script problem, and rewrite before scaling.

04

Replacement intent is not recognized at intake

The exact claim OBSERVED : In 5 designed calls presenting a competitor's written $8,500 full-system replacement quote, with an explicit offer to email the proposal on two of them, zero platforms asked to see the proposal, none asked what had been quoted, by whom, or when the decision would close, and soonest offered visits ran two to ten days where a date existed at all. One agent walked its own empty calendar month by month into September without escalating. One transfer to a human ended in under a minute.

0 of 5

platforms asked to see the competitor's proposal. Two were offered it by email, on the call, and declined. Exact 95% CI: 0% to 52%.

2–10

days to the soonest offered visit, where a date existed at all, for a buyer deciding that weekend.

15–20×

the average service ticket, riding on a call handled with the same queue, questions, and calendar as a $99 no-cool.

> "I was hoping to get an answer before Monday because I'm wanting to make up my mind this weekend about the purchase. So, I guess we're out of luck here."

ME, TO A PLUMBING BRAND'S AFTER-HOURS AGENT, ANNOUNCING THE CLOSE DATE OUT LOUD. THE AGENT OFFERED MONDAY, PLUS, TO ITS CREDIT, A SAVE: "WE'LL STILL SAVE YOUR SPOT FOR MONDAY, JUST IN CASE."

The denominator, stated bluntly. With five trials, a true ask rate as high as half the time is still consistent with observing zero. What the five calls can claim: five out of five independently sampled platforms, under roughly five different sponsors, ran the highest-value inbound call in the trade through the same process as a $99 ticket, and the behaviors that would distinguish it appeared zero times. That is a hypothesis with five supporting trials and no contradicting ones, and it is the single easiest hypothesis in this report to test in your own data.

External support EXTERNAL BENCHMARK : The speed literature is old, foundational, and directionally brutal. The Oldroyd/InsideSales analysis of 15,000-plus leads found qualification odds roughly 21 times higher at five-minute response versus thirty minutes; the 2011 HBR audit of 2,241 firms found firms responding within an hour were about seven times likelier to qualify a lead than those waiting longer, against an average response of 42 hours. Those studies measured web leads and B2B-heavy samples, so the multipliers do not transfer literally to a replacement-intent phone call; the decay shape is the transferable lesson, and a two-to-ten-day first touch sits far down every published decay curve. On value: the study's own price card puts a full replacement at $8,500 against service jobs of roughly $432, both quoted to me on these calls.

The alternative explanation. July capacity may genuinely have had no faster slots, and some brands may route second-opinion calls to comfort advisors through a backend process invisible to the caller. Both possible. Neither explains declining an emailed proposal that was offered mid-call, which costs no truck and no calendar slot.

Management implication INTERPRETATION : This chapter belongs to the platform whose paid search clears every month while booked replacement revenue stays flat. High-value intent should change the queue, the questions, and the clock. In this sample it changed none of the three; the one brand that learned my deadline learned it because I volunteered it, and still routed me into the standard flow.

CHECK THIS AT YOUR OWN PLATFORM MONDAY

- Pull 90 days of calls tagged replacement or estimate in ServiceTitan. Measure hours from call to first visit, by brand.
- Count what fraction of those records hold the competitor's quote as an attached document. Observed rate in this sample: zero of five.
- Listen to five of those calls and note whether anyone asks the close date. My sample's rate: zero, until the shopper volunteered it.

CONTROLLED PILOT · RECOGNIZE + ROUTE

In two markets, deploy replacement-intent routing: quote-in-hand callers skip the standard queue, land with an owner, trigger a same-day or next-day estimate slot, and the competitor quote gets captured as a document on the record. Compare against two matched markets on current practice for 60 days.

KPI: Replacement-intent speed to estimate (median hours, call to first appointment). **Guardrails:** Estimator utilization and standard-queue wait times, so expediting whales never starves the base business; estimate no-show rate. **Decision rule:** Scale if pilot markets' replacement booked-job rate rises against control with guardrails flat; if estimator capacity is the binding constraint, the finding converts to a capacity-planning decision with a known revenue number attached, which is exactly the decision a board can fund.

05

AI scales configuration, including configuration errors

The exact claim OBSERVED : Sixteen of 55 calls were fronted by a scripted AI agent; 13 ran end to end that way. Within those 16, the study observed the list every sponsor who signed the AI line item should read slowly:

- An agent gated scheduling behind an email address, then accepted one announced out loud as fake. The flow continued as designed.
- An agent had no waiver authority. Asked whether the trip fee would be waived on a purchase, it could only loop between the fee and a membership pitch. A human at a competing brand answered the same question by waiving the fee on the spot.
- An agent misvalidated my address, confirmed a town roughly 150 miles from the one I stated, and moved on. A truck dispatched on that record burns a slot and a customer.
- An agent looped through its own empty calendar, month after month into September, while a live replacement buyer waited, and never escalated to a human being.
- A greeting promised emergency availability; the agent answering it could neither price nor book one. The distance between those two is where the brand's credibility went.

"Sure, I will just give you, like, a fake email. Marcus 1234 at gmail dot com."
"Thanks, Marcus. And just to double check, about how old is your roof?"

EXCHANGE WITH AN AFTER-HOURS ROOFING AGENT. THE APPOINTMENT FLOW CONTINUED AS DESIGNED.

16 of 55

calls fronted by an AI agent; 13 ran end to end that way.

0

platforms in the sample could state the agent's call-to-booked-job rate on the human definition, on the observed evidence.

2

failed scheduling passes should trigger a human. One agent looped to September instead.

The governance frame EXTERNAL BENCHMARK , used as governance and never as a conversion benchmark: the NIST AI Risk Management Framework and its Generative AI Profile (NIST-AI-600-1, July 2024) organize AI risk work into four functions. Translated into controls for an inbound voice agent:

EXHIBIT 7 · AI GOVERNANCE FOR THE INBOUND VOICE CHANNEL

FUNCTION	CONTROLS FOR THE INBOUND VOICE AGENT
Govern	Written authority limits (what the agent may book, waive, promise); named configuration ownership across vendor, marketing, and ops; vendor accountability terms including audit access to agent logs; an incident log with severity definitions.
Map	An inventory of every flow the agent runs, every input it validates (address, contact, membership status), and every handoff point, with the observed failure modes mapped to specific flow steps.
Measure	The agent reported on call-to-booked-job on the human definition; validation-failure rates; escalation rates; post-change regression testing on a fixed call-scenario suite before any configuration ships.
Manage	Escalation to a human after two failed scheduling or validation passes; address validation against the CRM of record; a waiver table; monitored rollback paths; scheduled review of unbooked agent calls.

A necessary symmetry. This same sample's Finding 03 shows the agents doing the best retention selling observed anywhere in 55 calls. Both facts are configuration facts. Configuration executes identically on every call, which is why it compounds when right and scales the error when wrong. ServiceTitan and Genesys documentation describe the instrumentation this finding demands, subject to edition, license, and configuration.

Management implication INTERPRETATION : The AI budget is approved and spent; nobody needs to re-litigate it. The work the label promised and did not include is measurement and workflow. Pointing already-approved spend at instrumentation converts an experiment into an audited channel, in configuration work inside systems already owned.

CHECK THIS AT YOUR OWN PLATFORM MONDAY

- Ask each portco for the AI agent's call-to-booked-job rate on the human definition. Accept no substitute metric. The silence is the finding.
- Pull ten agent-handled recordings and score them against Exhibit 4. Fake-email gates, dead loops and mis-validations announce themselves inside the first five.
- Establish who can change the agent's configuration: vendor, marketing or ops. A channel only the vendor can tune is a channel the platform rents.

CONTROLLED PILOT · ROUTE + REVIEW

Trigger human escalation after two failed scheduling or validation attempts in half the agent's traffic (or alternating weeks), against current behavior in the other half. **KPI:** Agent call-to-booked-job on the human definition. **Guardrails:** Escalation volume landing on human queues; false-escalation rate; caller abandonment during handoff. **Decision rule:** Scale the escalation rule if agent-path booked-job rate rises without swamping human queues; if escalations swamp queues, the finding is an understaffing number, priced and fundable.

MODELLED, THROUGHOUT

Illustrative economic exposure

The prior edition of this study published a headline of $1.8M to $2.1M of install EBITDA "at risk" from the replacement-lead leak at 50,000 annual demand calls. That figure was the full expected pipeline riding on the observed handling, and full pipeline is the wrong loss estimate: mishandling does not zero a close rate. This section replaces it with a scenario model built on one formula, with every input sourced or flagged:

Incremental EBITDA opportunity = eligible demand volume × average economic value × estimated incremental conversion lift × contribution margin − implementation cost

EXHIBIT 8 · WORKED MODEL: REPLACEMENT-INTENT HANDLING, 50,000 DEMAND CALLS A YEAR

INPUT	VALUE	LABEL, AND WHAT REPLACES IT
Demand calls per year	50,000	Company input, illustrative. Replace with your CCaaS offered-call count, marketing lines only.
Replacement-intent share	5%	ASSUMPTION My five calls were a designed test and measure no incidence; placeholder pending your ServiceTitan call reasons tagged replacement or estimate.
Average economic value	$8,500	OBSERVED The competitor quote carried on these calls; conservative against quoted replacement ranges. Replace with your average install ticket.
Incremental conversion lift	3 / 7 / 12 pts	ASSUMPTION Directionally supported by the response-speed literature, which measured different channels; no home-services experimental estimate exists in the cited sources. Your pilot produces this number in 60 days.
Contribution margin	25 / 27.5 / 30%	ASSUMPTION Deliberately below typical trade gross margins; excludes incremental marketing by construction.
Implementation cost, year one	$40K / $75K / $120K	ASSUMPTION Configuration labor, QA time, training, possible estimator overtime; no software purchase assumed.

EXHIBIT 8B · SCENARIOS, SENSITIVITY, AND RECOVERABILITY

SCENARIO (2,500 ELIGIBLE CALLS; $21.25M ELIGIBLE PIPELINE)	REVENUE LIFT	EBITDA BEFORE COST	NET OF IMPLEMENTATION
Low (3 pts, 25%)	$637K	$159K	~$40K to $119K
Base (7 pts, 27.5%)	$1.49M	$409K	~$290K to $335K
High (12 pts, 30%)	$2.55M	$765K	~$645K to $725K

SENSITIVITY: EBITDA BEFORE COST, $K	25% MARGIN	27.5%	30%
3-point lift	159	175	191
5-point lift	266	292	319
7-point lift	372	409	446
10-point lift	531	584	638
12-point lift	638	701	765

Recoverability, explicitly. The lift assumption already embeds recoverability: it prices only the incremental installs the improved handling wins, never the full pipeline. Even the high case claims well under half the gap between observed handling and a perfect close, and the low case still clears its own implementation cost. **The other four leaks** (dead-ended calls, price opacity, unrecovered deferrals, unattached second problems) run through the same formula with smaller tickets and are individually smaller under any defensible assumption set. **Do not sum the leaks:** the populations overlap, and a summed figure double-counts. Rank them, pilot the top one, and let measured lift replace modeled lift line by line. Your call-reason data replaces the 5 percent; your pilot replaces the lift; your finance team replaces the margin; and once those three land, the model stops being mine.

Six implementation cards

Every card lives inside systems the sampled platforms already run; edition and configuration dependencies are flagged where the vendor documentation flags them. Each card names an executive owner, an operational owner, a system owner, a pilot, a KPI, guardrails, the principal risk, and a scale decision.

CARD 1 · WRITTEN RECAP AT CALL WRAP · RETAIN

Problem: 1/55 recaps; 36/49 answered calls never asked for an email. **Intervention:** Email as a required intake field with a readback step; SMS or email recap fired at wrap from ServiceTitan or the CCaaS. **Owners:** CRO / CSR lead per brand / ServiceTitan admin. **Configuration:** Required intake field; recap template; wrap-trigger automation. **Dependencies:** Sending enabled; consent language. **Effort:** Days. **Pilot:** Randomize eligible unbooked calls into recap and control. **KPI:** Unbooked-call recovery at 14 days. **Guardrails:** Handle time; opt-out rate; fake-contact rate. **Principal risk:** Forced capture generates junk data and caller friction. **Mitigation:** Readback step; permit "no email" with a reason code. **Scale decision:** Recap arm recovers 3+ points more unbooked calls over 8 weeks, guardrails flat.

CARD 2 · THE SAVE SCRIPT, DEPLOYED AT DAYTIME VOLUME · RETAIN

Problem: 2/33 human save attempts against 6/13 for agents. **Intervention:** The agent's configured save line, taught to the daytime team in one morning huddle with a wallet card. **Owners:** CRO / CSR lead / QA sampling in the CCaaS. **Effort:** One morning plus weekly QA. **Pilot, KPI, guardrails, decision rule:** Finding 03's card.

CARD 3 · REPLACEMENT-INTENT ROUTING · RECOGNIZE + ROUTE

Problem: 0/5 proposals captured; visits 2 to 10 days out. **Intervention:** Quote-in-hand callers skip the standard queue, land with an owner, trigger a same-day estimate slot; quote captured as a document. **Owners:** COO / intake lead / CCaaS admin + ServiceTitan admin. **Dependencies:** Estimator capacity; routing configurability. **Effort:** Weeks. **Pilot, KPI, guardrails, decision rule:** Finding 04's card.

CARD 4 · ABANDONED AND UNBOOKED CALL RECOVERY · RETAIN + REVIEW

Problem: 6/55 calls dead-ended at brands running live paid media; 11 promised callbacks, 2 observed arrivals. **Intervention:** Timed callback workflow on abandoned and unbooked calls; a "Callback promised" call reason required at wrap, auto-generating a same-day task; a daily report of promised callbacks with no booked job inside 24 hours. ServiceTitan documents call reasons, task workflows, and abandoned-call reporting; Genesys documents abandon intervals and short-abandon exclusions for clean denominators. **Owners:** COO / CSR lead / ServiceTitan + CCaaS admins. **Effort:** A configuration afternoon to start measuring. **Pilot:** Timed callbacks versus current practice, alternating weeks. **KPI:** Callback-kept rate; abandoned-call recovery rate. **Guardrails:** Outbound attempts per CSR; complaint rate. **Principal risk:** The ledger exposes uncomfortable numbers and gets quietly dropped. **Mitigation:** The past-due callback count goes on the weekly ops review, by name. **Scale decision:** Kept rate above 80% sustained for a month, then extend to unbooked-call re-dials.

CARD 5 · SIBLING REFERRAL MACHINERY · ROUTE

Problem: 0/55 referrals. **Intervention:** A transfer path and a lead-source code per sibling brand, so the zero has somewhere to move. **Owners:** Operating partner (cross-brand by definition) / intake leads / CCaaS + CRM admins. **Dependencies:** Service-area and licensing checks; cross-brand revenue credit rules agreed before launch, because unresolved credit kills referrals faster than any technical gap. **Effort:** Weeks. **Pilot, KPI, guardrails, decision rule:** Finding 01's card.

CARD 6 · AI AGENT GOVERNANCE · ROUTE + REVIEW

Problem: Finding 05's failure list; no agent booked-job number anywhere. **Intervention:** The NIST-mapped control set run as configuration work: escalation after two failed passes, address validation against the CRM, a waiver table, a fixed regression suite after every configuration change, the agent reported on call-to-booked-job like any CSR. **Owners:** COO / intake lead / whoever the Finding 05 diagnostic reveals actually controls configuration; if the answer is "the vendor," the first deliverable is contractual audit and change access. **Effort:** Weeks. **Pilot, KPI, guardrails, decision rule:** Finding 05's card.

EXHIBIT 9 · THE 30-, 60-, 90-DAY PLAN

Measure, pilot, scale. In that order.

Days 1 to 30: measure before touching

Write the metric dictionary (next section's definitions, adopted verbatim or amended in writing). Validate the data: reconcile CCaaS offered-call counts against ServiceTitan call records for one week, and resolve gaps before trusting any rate. Pull baselines by brand and daypart. Run a 50-call listening sample scored on the eight stages. Name the executive, operational, and system owner for each card. Select two pilots, no more: the recap randomization plus one of Cards 3, 4, or 6, chosen by the baseline.

Days 31 to 60: pilot under daily watch

Launch both pilots with control groups as designed. Daily monitoring of primary KPI and guardrails; weekly QA on both agent-handled and human-handled calls in the pilot population; correct configuration issues within 48 hours of detection and log every change. The regression suite from Card 6 applies to human-facing configuration changes too.

Days 61 to 90: scale what measured, standardize what scaled

Scale pilots that hit their decision rules; kill or redesign the rest, in writing. Transplant the winning configuration platform-wide (the Finding 02 mechanism). Stand up the board dashboard with the first month of real data. Embed the cadence: weekly ops review carries the callback ledger and abandonment; monthly QA carries the rubric scores; the board pack carries call-to-booked-job next to marketing spend.

EXHIBIT 10 · THE BOARD DASHBOARD · METRIC DICTIONARY

Thirteen metrics: numerator, denominator, exclusions, window, source, owner, cadence.

Where brands differ on any definition today, the first deliverable is this table signed by every brand president.

METRIC	NUMERATOR / DENOMINATOR	EXCLUSIONS · WINDOW	SOURCE · OWNER · CADENCE
1 Qualified inbound demand calls	Calls with lead call reasons / all inbound	Internal, vendor, spam, existing-job status · same-day	CCaaS + ST call reasons · CSR lead · weekly
2 Answer rate	Answered by live agent or booking-capable AI / offered demand calls	Short abandons under configured threshold · same-day	CCaaS · COO · weekly
3 True abandonment	Abandoned after short-abandon threshold / offered demand calls	Short abandons, per Genesys definitions · same-day	CCaaS · COO · weekly
4 Call-to-booked-job	Booked jobs from qualified demand calls / qualified demand calls	Non-lead reasons; excused calls audited monthly · 48-hr attribution	ST Calls dataset · CRO · weekly, board monthly
5 Booking cancellation rate	Booked jobs canceled before dispatch / booked jobs	Customer reschedules kept within 7 days · 14-day	ST · COO · monthly
6 Completed-job rate	Completed jobs / booked jobs	As metric 5 · 30-day	ST · COO · monthly
7 Replacement speed to estimate	Median hours, call to first appointment, replacement-tagged	Customer-requested delays, logged with reason code · per call	ST · COO · weekly
8 Quote-capture rate	Replacement calls with competitor quote attached / replacement calls where a quote was disclosed	None · same-day	ST · CSR lead · monthly
9 Unbooked-call recovery	Unbooked demand calls with a booked job inside 14 days / unbooked demand calls	Known competitor bookings (rarely known; note the bias) · 14-day	ST + task reports · CSR lead · monthly
10 Sibling referral eligibility & completion	Completed sibling referrals / calls with a voiced second or out-of-trade need	Eligibility from QA sampling until a call reason exists · 30-day	CCaaS transfer logs + sibling lead-source codes · operating partner · monthly
11 Membership attach	Memberships sold / membership-eligible calls, eligibility written per brand	Existing members · same-day	ST · brand president · monthly
12 AI vs human outcomes	Metrics 2, 4, 5, 6 split by answering channel, same definitions	As parent metrics	CCaaS agent reporting + ST · COO · monthly
13 By brand and daypart	Metrics 2 to 4 cut by brand and business/after-hours daypart	As parent metrics	CCaaS · COO · monthly

THE MONDAY-MORNING SELF-TEST

Ten questions, formulas unredacted

The first seven cost one hour: pull five recorded calls per brand and listen. The last three are data pulls, formulas published in full. Score yes or no; no partial credit. Then score each of the five controls on the maturity scale: **1 Absent · 2 Inconsistent · 3 Standardized · 4 Measured · 5 Managed against target.** For calibration: on the listening evidence, the median brand in this sample sits at 1 to 2 on Retain and Route, 2 to 3 on Reach, and Review is unobservable from outside, which is itself the Review finding.

Listen · one hour

1. Does anyone ask for the caller's email address? (This sample: 13 of 49.)
2. When the caller defers ("I'll talk to my spouse"), is the next sentence a save attempt, or "no worries"? (8 of 49 attempted.)
3. Can the person answering state the price of a visit, and whether it credits against the work? (9 of 49 could not.)
4. When a second problem is mentioned mid-call, does it get one trip, a second fee, or "call someone else"? (3 of 13 got one trip.)
5. Does anything arrive in writing after the call: text, email, quote? (1 of 55.)
6. When the caller holds a competitor's quote, does anyone ask to see it? (0 of 5.)
7. Call your own after-hours line tonight. Can the voice that answers quote the fee, book a job, or commit to a callback with a time on it? (This sample: mostly no.)

Pull · three reports, full formulas

1. **Call-to-booked-job, by brand:** booked jobs created within 48 hours of a qualified demand call ÷ qualified demand calls, excluding non-lead call reasons and audited excused calls, one definition across every brand. If the definitions differ by brand, that finding outranks the number.
2. **Speed to answer and true abandonment, by brand and daypart:** abandoned calls after the short-abandon threshold ÷ offered demand calls, with short abandons excluded before the rate means anything; average speed of answer computed on answered calls only.
3. **Membership attach:** memberships sold ÷ membership-eligible calls, where eligibility is written down per brand and existing members are excluded from the denominator.

If the seven listening checks come back clean at your platform, close this report and keep the hour. In this sample of 48 brands, none would have come back clean.

RISKS AND TRADE-OFFS

Every fix above has a failure mode

Forced contact capture breeds fake data and caller friction; Finding 05's fake-email exchange shows validation-free capture is worse than none. Mitigate with readback, validation, and a "declined" reason code. **Consent and privacy:** SMS recaps and recorded-call QA run into consent requirements that vary by state and province; legal review precedes Card 1, and this report is not legal advice. **Estimator capacity:** replacement-intent routing without capacity planning converts a routing win into a no-show factory; the guardrails in Finding 04 exist for this.

Low-quality bookings and cancellation risk: every save script and booking push can inflate booked jobs that never complete, which is why cancellation and completed-job rates sit beside booking rate on the dashboard, and why no pilot in this report scales on attempted behaviors alone. **Local brand autonomy:** standardization has real costs in markets where the local brand's voice is the asset; standardize the controls (capture, recap, routing, measurement) before touching brand voice. **Cross-brand revenue credit:** unresolved credit allocation quietly kills referral machinery; agree the split before the pilot, in writing. **Service-area and licensing limits** legitimately constrain routing across brands and state lines; map them in the Finding 01 pilot design.

Vendor control over configuration: a channel only the vendor can tune is a channel the platform rents; contractual audit and change access is a deliverable. **Inconsistent data definitions** are the quiet killer of every metric on the dashboard, which is why the signed metric dictionary is deliverable one. **AI failure escalation** can swamp human queues if thresholds are set naively; stage the rollout. **Employee adoption:** config compounds, coaching decays, and resentment compounds faster than either. The two human saves in this sample were the best selling observed anywhere, so the message to CSR teams is that the machinery exists to route them better at-bats, and the QA cadence exists to find coaching moments rather than culprits.

CONCLUSION

The strongest defensible conclusion, and the decision that follows

Across 55 calls to 48 PE-backed brands, the intake behaviors that turn answered calls into booked revenue (the save, the capture, the recap, the kept follow-up, the sibling route, the recognition of replacement intent) were observed rarely, inconsistently, or never, while the software layer those platforms already pay for mostly worked, and the scripted layer, where deployed, executed its configured selling behavior far more consistently than the human layer ($p = 0.004$). The external record says the missing behaviors are the ones that decide conversion, the missing synergy class is the one that decides the deal model, and the missing measurements are available in the platforms' own documented systems.

The decision executives should make next is small and this quarter: adopt one metric definition for call-to-booked-job across every brand, run the ten-question self-test, and fund one controlled pilot with a written decision rule. Paid demand becomes measurable revenue only when the operating layer is controlled, and the operating layer is controlled only when someone owns it, measures it, and reviews it. In this sample, on the evidence of 55 calls, nobody did. **The front door is the integration test. Most platforms in this sample have not yet taken it, and the caller can tell.**

External sources

Graded per the report's source hierarchy: government and standards bodies; official product documentation; empirical research; recognized professional-services research; vendor research with disclosed data. Thirteen of sixteen sources sit in the top categories; vendor research is flagged as vendor research. Deliberately excluded: SEO aggregator statistics, anonymous conversion claims, and vendor assertions without disclosed samples.

- NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023. Government standards body.
- NIST, *AI RMF: Generative Artificial Intelligence Profile*, NIST-AI-600-1, July 26, 2024. Government standards body.
- ServiceTitan, "Data Report: Average Call Booking Rates," October 2022. Vendor data, June 2022 platform cross-section, all trades: 42% average booking rate; 59% at 25+ technicians; 24% under 5.
- ServiceTitan Help Center, Contact Center Pro documentation set: product home; voice-agent settings; routing integration; queue monitoring; Agent Performance report definitions; call reasons and lead classification; job-booking workflows. Accessed July 2026. Official product documentation; capabilities flagged as Pro-product, edition and configuration dependent.
- ServiceTitan, "Spring into Action: Using ServiceTitan's Benchmark Report," May 2024. Vendor guidance on call-reason hygiene, excused-call auditing, permissions.
- Genesys Cloud Resource Center, "Service level" glossary (abandon inclusion and exclusion, short-abandon threshold, numerator and denominator options), updated November 2025; "Abandon Intervals Metrics view," January 2026; workforce management metric definitions, 2024. Official product documentation.
- Invoca, *Call Conversion Industry Benchmarks Report 2025*, June 2025. Vendor research, 60M+ anonymized calls on Invoca's customer base; self-selection noted; conversion defined as signal during call. Figures used: 61% reach a person; answer rates 54–69% by industry; 37% cross-industry lead conversion during call; home services 46%; 35% of agents ask for purchase or booking.
- Kroll, *M&A in the Residential HVAC Services Industry*, November 2025. Professional-services research.
- Capstone Partners, *HVAC Services M&A Update*, July 2025. Professional-services research; add-on activity +88.2% year over year, deal-count tracking with disclosed method.
- McKinsey & Company, "Seven rules to crack the code on revenue synergies in M&A," 2018. Survey of 200 M&A executives; 23% average shortfall; capture-timeline analysis.
- McKinsey & Company, "Capturing cross-selling synergies in M&A," February 2020. Cross-selling at roughly 20% of revenue-synergy value.
- McKinsey & Company, "Focusing your M&A team on revenue growth," 2017. Global-1000 analysis; seven-percentage-point sales-growth penalty.
- Oldroyd, J., with InsideSales.com, *Lead Response Management Study*, 2007. Six companies, 15,000+ leads, 100,000+ call attempts; 100× contact and 21× qualification odds at 5 versus 30 minutes. Foundational and pre-2024, retained per source policy; web-lead channel difference disclosed wherever cited.
- Oldroyd, J., McElheran, K., Elkington, D., "The Short Life of Online Sales Leads," *Harvard Business Review*, March 2011. Audit of 2,241 US firms; 42-hour average response; 37% within an hour; roughly 7× within-hour qualification.

Companion supplement, available on request: the full evidence matrix (claim by claim: field evidence, denominator, external support, strength, alternative interpretation, permitted wording, remaining gap), statistical notes, economic-model inputs, exhibit specifications, the outstanding-data register, and a recommended replication design. zaha@ztsadvisory.com.

ABOUT THE AUTHOR

I run revenue-capture work inside PE-backed residential service platforms, hands-on in the systems this report names: ServiceTitan, HubSpot, and the CCaaS layer on top of them. This study points the same method I use inside platforms outward: blind mystery-shop calls, one fixed rubric, and the discipline of separating what was observed from what is modeled. The report speaks for itself; so should yours. Pull the ten questions and find out.

IF THE SELF-TEST COMES BACK THE WAY IT DID IN THIS SAMPLE

The Revenue Leak Read

ZTS runs a two-week, fixed-fee Revenue Leak Read. It runs the three pulls (call-to-booked-job by brand; speed to answer and true abandonment by brand and daypart; membership attach by brand) on the platform's own ServiceTitan and CCaaS data, listens to a structured sample of recorded calls scored on the eight stages and the five controls, and returns a board-ready read-out with the exposure modeled on the platform's own numbers, observed and modeled kept as strictly separate as they are in this report. It closes with a prioritized fix list, mapped to the six implementation cards, that your own team can run.

TWO WEEKS · FIXED FEE · SELF-CONTAINED · NO SOFTWARE TO BUY

The read is complete in itself. Nothing else is attached to it.