

CUSTOMER EXPERIENCE DESIGN & INNOVATION

The Agentic CX Performance System

Containment is not resolution. Usage is not adoption. Automation is not value. This is how to measure what an agent actually delivers — and when it has earned more room.



Three substitutions that hide failure

Volume I made the case for designing agentic systems. Volume II built one. This volume answers the question both leave open: *we launched it — how do we know it is working?* Most programs measure something next to the truth, and three substitutions explain why.

FIGURE 1 • WHAT GETS REPORTED, AND WHAT IT IS MISTAKEN FOR

WHAT THE DASHBOARD SAYS

Containment

The conversation ended without a person.

Usage

Customers passed through the agent.

Automation

Work moved from people to software.

WHAT THE BOARD HEARS

Resolution

The need was met, confirmed in the record, and did not come back.

Adoption

Customers chose it again when a person was one click away.

Value

Cost to serve, retention or revenue moved — against a holdout.

Each left-hand number is real, cheap to collect and usually rising. Each right-hand number is what leadership believes it is being shown. The distance between the two columns is where agentic programs lose credibility.

What this paper is for

For the executives who fund, own and answer for customer-facing agents — CEO, CFO, CMO, CPO and CX leadership. It gives them a hierarchy of measures, a value model linking agent performance to retention, revenue and cost to serve, a live-agent scorecard, a benefits-realization template, and written rules for widening — or narrowing — an agent's autonomy.

What it is not

Not an analytics architecture or a dashboard comparison. Every measure here can be built from logs most programs already keep. The constraint is agreeing, before launch, what counts as success.

Illustrative figures continue the Volume II worked build — a retailer's order-resolution agent — one quarter after launch.

14%

of service issues fully resolved in self-service

GARTNER, 2024

7 in 10

agent conversations end without escalation

SALESFORCE, 2026

25%

of AI initiatives delivered their expected ROI

IBM, 2025

21%

of brands' CX quality declined in 2025; 6% improved

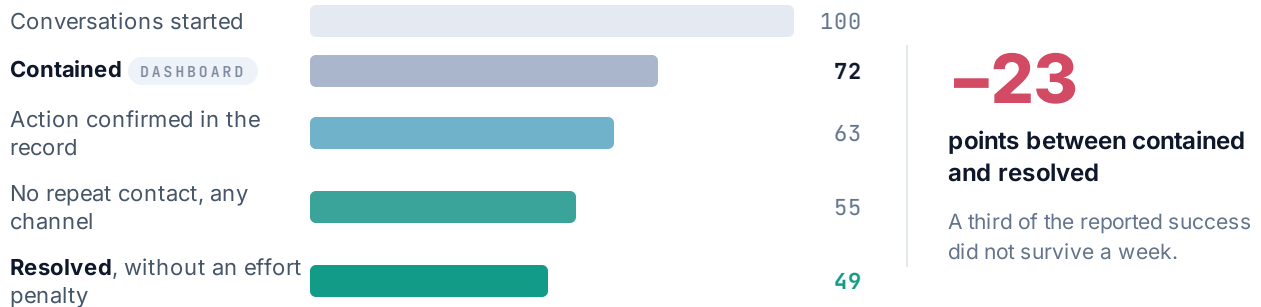
FORRESTER, 2025

The first two numbers measure different things, in different years — one counts conversations that did not escalate, the other asks customers whether the problem was solved. That difference is the subject of this paper.

Green dashboard, returning customer

Containment counts conversations that ended without a person. It cannot see the customer who gave up, phoned instead, or returned on Tuesday with the same problem. Follow one hundred conversations past the point where the dashboard stops, and the number changes shape.

FIGURE 2 • ONE HUNDRED CONVERSATIONS, FOLLOWED FOR SEVEN DAYS



Illustrative, from the Volume II build. The gap sits in two places: conversations that ended before the action completed, and customers who returned through a channel the agent's dashboard cannot see.

Why the gap stays invisible

Containment is measured inside the channel. Failure shows up outside it. A customer who abandons chat and phones the contact center appears in one report as a success and in another as a new call — and nobody joins the two.

Gartner found only 14% of service issues fully resolve in self-service, and 45% of customers who started there said the company did not understand what they wanted. Agents resolve far more. The measurement habit carried over anyway: count exits, not outcomes.

WORTH NOTICING

The gap is usually widest where the dashboard looks best. Containment above its healthy band is more often a trapped customer than a capable agent.

The fix is a definition, not a tool

Resolution is an action confirmed in a system of record, and no contact on the same intent within seven days, in any channel. Everything later in this paper depends on agreeing that sentence first.

SUBSTITUTION ONE

Containment ≠ resolution

Resolution is confirmed in the record and stays solved for seven days. Ending the chat is not evidence of either.

SUBSTITUTION TWO

Usage ≠ adoption

Forced routing produces usage. Adoption is choice — customers who could reach a person and came back to the agent anyway.

SUBSTITUTION THREE

Automation ≠ value

Value is a difference against a holdout in cost, retention or revenue. Work that moved without moving any of those is a cost shift.

Seven measures, each earning the next

The fix is not more metrics. It is order. Each measure below is credible only if the one beneath it holds — a cost figure means nothing until resolution is proven, and resolution means nothing until the task actually completed.

FIGURE 3 • THE HIERARCHY OF AGENT PERFORMANCE

	MEASURE	THE QUESTION IT ANSWERS	GAMED BY, IF READ ALONE
ECONOMICS	07 Financial return	Did retention, revenue or cost to serve move — against a holdout?	Crediting a market trend
	06 Cost per resolved intent	What did a real resolution cost, all in?	Dividing by conversations
EXPERIENCE	05 Trust	Do customers choose to delegate again?	Hiding the way out
	04 Customer effort	How hard was it to get there?	Surveying only finishers
RESOLUTION	03 Handoff quality	When a person took over, did the context arrive?	Suppressing escalation
	02 Repeat contact	Did the customer have to come back, in any channel?	Same-channel matching
	01 Task completion	Is the action confirmed in the system of record?	Counting ended chats

Read up to report, down to diagnose. The board sees level seven. When it moves the wrong way, the cause is almost always two to five levels below it — and a program that only instruments the top cannot find it.

Why effort sits in the middle

Effort is the bridge between operations and loyalty. CEB research, now Gartner's, found 96% of customers with high-effort service interactions became more disloyal, against 9% with low-effort ones.

An agent that completes the task in eleven turns and two clarifications has done the job and spent the relationship.

Why trust is measured as behavior

Surveys ask whether customers trust the agent. Behavior answers: do they undo what it did (reversion), leave at the first turn (opt-out), or return to it when a person was one click away?

That last one — the **return-to-agent rate** — is the closest thing to adoption you can measure.

An agent is not working because it is busy. It is working when customers stop needing to come back.

The unit of account is the resolved intent

Conversations are what the platform counts. Intents are what the customer came for. Every measure below is defined against the intent, over a stated window, with the trap that most often breaks it.

FIGURE 4 • DEFINITIONS THAT HOLD UP IN A BOARD MEETING

MEASURE	COUNTS AS SUCCESS WHEN	WINDOW	THE TRAP
Task completion	The action is confirmed in the system of record and a reference is returned	Session	Treating “conversation ended” as done
Repeat contact	No further contact on the same intent, in any channel	7 days	Matching only inside the agent's channel
Handoff quality	The person receives intent, history and attempted actions; the customer repeats nothing	Per handoff	Timing the transfer instead of auditing the context
Customer effort	Few turns, little elapsed time, no channel switch to reach resolution	Per intent	Asking only customers who completed
Trust	Low reversion, stable first-turn opt-out, rising return-to-agent rate	30 days	Reading low opt-out as trust when the exit is hidden
Cost per resolved intent	All-in cost ÷ intents resolved — not conversations handled	Monthly	Leaving repeat contacts and rework out of the numerator
Financial return	Difference against a holdout in churn, revenue and cost to serve	Quarterly	Before-and-after in a moving market

FIGURE 5 • THE DENOMINATOR CHANGES THE STORY



Gartner expects generative AI cost per resolution to exceed \$3 by 2030, surpassing many offshore human agent costs. A case built on cost per conversation can cross that line without anyone noticing, because the denominator hides it.

THE RULE

Repeat contacts belong on the agent's side of the ledger. If the agent's miss is handled by a person, the person's cost is the agent's cost.

What moves this week, what pays next year

Boards want lagging outcomes: retention, revenue, cost to serve. Operators can only act on leading ones. The discipline is knowing which leading indicator predicts which outcome, how long the lag is — and never reporting one as the other.

FIGURE 6 • THREE CLOCKS, FOUR CHAINS



Lags are typical ranges from the worked build, not laws. Measure your own. The lag between a leading indicator and its outcome is itself a finding — and it tells the board when to expect proof.

Manage to leading, report on lagging

The build team should never wait for churn to move. The board should never be told containment rose as if that were a result. Each audience gets the clock it can act on.

- **Reporting a leading indicator as an outcome.** “Containment rose” is activity, not a result.
- **Waiting on a lagging outcome to act.** By the time churn moves, a quarter of customers has already had the bad version.

THE TEST

For every leading indicator on your scorecard, name the lagging outcome it predicts and the lag. If you cannot, it is activity data, not a performance measure.

Put the lag on the page

Show each lagging outcome with the date its evidence is due. It ends the quarterly question — “why hasn't churn moved yet?” — before anyone asks it.

+10

NPS points on a named concierge deployment

SALESFORCE, 2026

+24%

first-contact resolution in scaled programs

MCKINSEY, 2026

70%

see measurable value within 60 days

SALESFORCE, 2026

\$3+

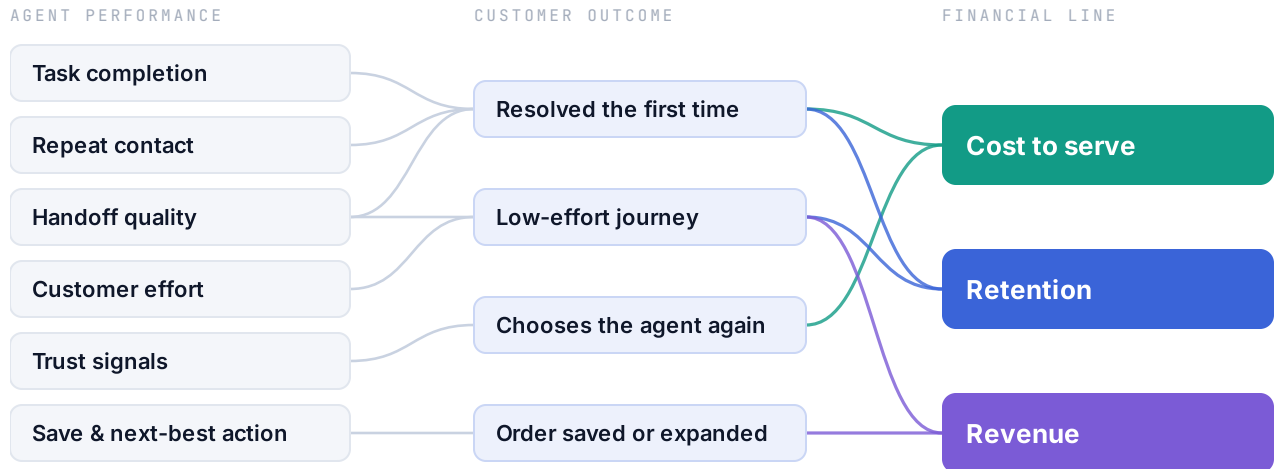
expected AI cost per resolution by 2030

GARTNER, 2026

Three lines to the P&L

Agent performance reaches the P&L through exactly three lines: cost to serve, retention and revenue. Most business cases model the first, assume the second and ignore the third. The model below connects each line to the measures that drive it.

FIGURE 7 • FROM AGENT PERFORMANCE TO FINANCIAL OUTCOME



Resolution feeds two lines, not one. A first-time resolution lowers cost to serve and protects retention — which is why a cost-only business case undercounts the agent that works, and overcounts the one that deflects.

COST TO SERVE

resolved intents × cost per resolved intent + escalations × assisted cost

Repeat contacts sit inside "cost per resolved intent," not beside it.

RETENTION

(holdout churn – served churn) × customers served × lifetime value

Only a holdout makes this line defensible.

REVENUE

orders saved or expanded × margin – what the holdout saved anyway

The line most cases leave at zero.

Why retention is the line that matters

Cost savings are linear and visible. Retention effects are small, delayed and compounding — so they lose every budget argument they are not modeled into. Model them explicitly, with a holdout, or the agent will be managed as a cost program forever.

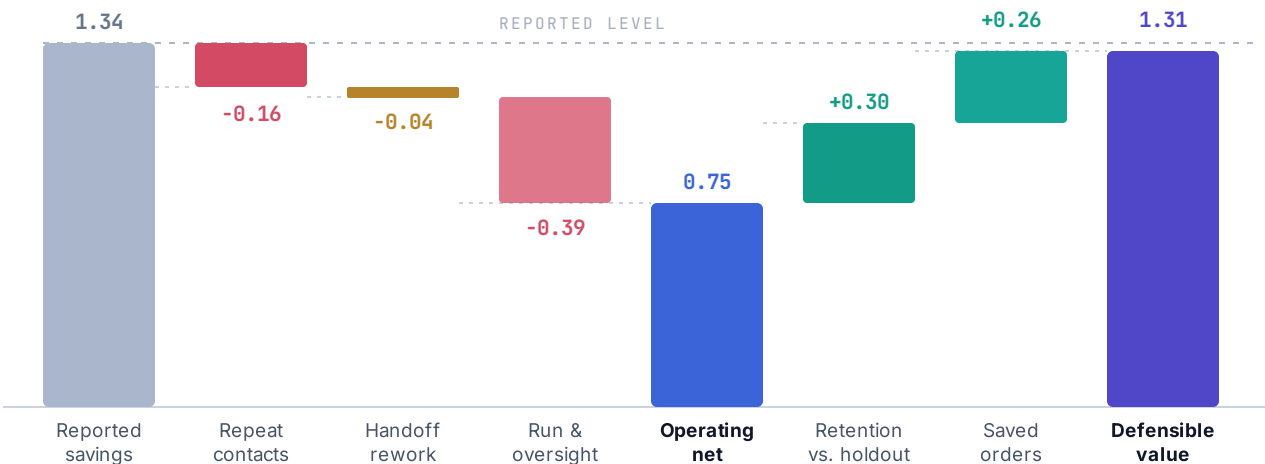
A NOTE ON THE FAMOUS NUMBER

Bain's Frederick Reichheld found a 5% rise in retention could lift profits 25% to 95%. It is old, sector-specific and widely over-quoted. Use its shape — small retention moves outweigh large cost moves — not its range.

The claimed number and the defensible one

One quarter after launch, the Volume II retailer's agent reported \$1.34 million a year in automation savings. Rebuilt with the value model, the total barely moves. What it is made of changes completely.

FIGURE 8 • ONE QUARTER AFTER LAUNCH, ANNUALIZED (\$M)



The headline and the net land close together — for entirely different reasons. 44% of the reported savings did not exist. 43% of the real value sat in two lines nobody was reporting. Illustrative, 364,800 intents a year.

Why the reasons matter more than the total

A CFO who approves \$1.34M is approving a number the next audit of repeat contacts will falsify. A CFO shown \$1.31M built this way approves a number that survives one — and learns the agent's largest upside is retention and saved orders.

That changes what gets funded next.

Hold out, or you are guessing

Before-and-after comparisons credit the agent with whatever else changed that quarter. Three designs hold up, in order of strength.

THE QUESTION THAT EXPOSES A WEAK CASE

"Compared with what?" If the answer is last year, the case is a trend line with a logo on it.

STRONGEST

Holdout

A fixed 5–10% of eligible customers stays on the prior experience. It costs a little resolution to buy certainty.

PRACTICAL

Staggered rollout

Journeys or regions switch on in sequence. The not-yet-launched group is the comparison.

LAST RESORT

Matched cohort

Served customers compared with statistically similar ones who were not. Use when the other two are impossible.

One page, four rows of evidence, one decision

Volume II's scorecard proved a pilot. This one runs a live agent: it adds the measures that only matter at scale, and ends every review with a recorded decision.



WEEKLY • BUILD TEAM

Levels one to three; every trip and error

MONTHLY • OWNER

Full page; one decision per intent

QUARTERLY • LEADERSHIP

Value vs. holdout; benefits realized

TWICE A YEAR • BOARD

Net value, trust trend, autonomy changes

Sign the benefits, then read them

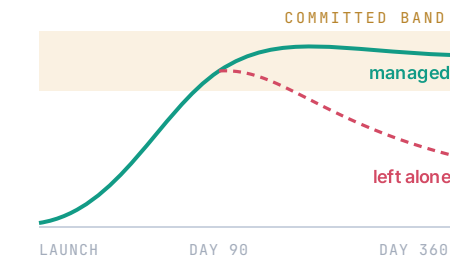
Most AI business cases secure funding and are never opened again. This template turns the case into a contract: each benefit gets a baseline, a committed band, an attribution method and an owner — agreed before launch.

FIGURE 10 • BENEFITS-REALIZATION TEMPLATE, FILLED FOR THE WORKED BUILD

BENEFIT • MEASURE	BASELINE	COMMITTED BAND	ATTRIBUTION	OWNER	DAY 30	DAY 90	STATUS
Lower cost to serve cost per resolved intent	\$5.10	\$2.20–\$3.00	Holdout	Dir. Customer Ops	\$3.40	\$2.57	ON TRACK
Fewer repeat contacts 7-day, any channel	11%	7–9%	Holdout	CX Ops lead	12%	8.9%	ON TRACK
Retained customers 90-day churn vs. holdout	6.2%	–0.4 to –0.8 pts	Holdout	VP Customer Experience	—	–0.6 pts	EARLY READ
Saved orders cancel-to-exchange rate	1.1%	3–5%	Matched cohort	eCommerce GM	2.0%	4.0%	ON TRACK
Lower effort journey effort score	5.1	≥ 5.5	Pre/post, same journey	Design lead	5.3	5.8	MET
Handoff quality context complete	—	≥ 90%	Audit sample	Service lead	78%	84%	BEHIND
Dis-benefit complaints about agent actions, per 10k	3.2	≤ 3.2	Holdout	Risk partner	3.6	3.0	HELD

The last row is the one most templates leave out. Every agent creates some new failure. Naming it in advance, with an owner, is what makes the rest of the template credible to a CFO.

FIGURE 11 • REALIZATION CURVE



Benefits decay when knowledge ages, policy changes and nobody reads the page.

Five rules for a case that survives audit

- Signed before launch** by a business owner — not the vendor or the build team.
- Bands, not points.** A point invites gaming; a band shows both directions.
- Attribution named in advance.** Choosing the method after seeing the data is choosing the answer.
- Dis-benefits included** — new complaint types, volume pushed elsewhere, knowledge upkeep.
- Read on a date, not on request.** Day 30, 90 and 180, in calendars before launch.

Evidence widens. Evidence narrows.

Volume I argued that autonomy is granted per moment and raised on evidence. These are the rules, written down. Each applies to one intent at one tier — and the evidence that widens authority must be able to narrow it, or the rules are a ratchet.

FIGURE 12 • THE AUTONOMY DECISION RULES

EXPAND	HOLD	REDUCE	SUSPEND
One tier, one intent. All hold for 8 weeks or 500 resolved intents: - Task completion $\geq 95\%$ - Repeat contact at or below the human baseline - Reversion $< 2\%$ - Calibration $> 80\%$ - Confirm-first approvals $\geq 97\%$, with real review time - No severe incident; trust flat or rising	Stay at the current tier when any apply: - The sample is too small to decide - A measure is in band but moving the wrong way three weeks running - A model, prompt, knowledge or tool change shipped in the window — the clock restarts	Down one tier, the same day, when any apply: - Repeat contact above the human baseline two weeks running - Reversion $\geq 5\%$ - Calibration $< 70\%$ - Complaints rising about one action type - Handoff context $< 80\%$	To Tier 3 — escalate only — immediately, on any one: - Money moved beyond granted authority - A disclosure or consent failure - Harm in an excluded moment - A guardrail bypass, or a stated intent that violates policy

One intent, one tier, one change per window. Thresholds are the worked build's; set yours from your own baseline. Write them before launch, change them only by version, and record every decision against the role card in force.

Widen slowly, narrow fast

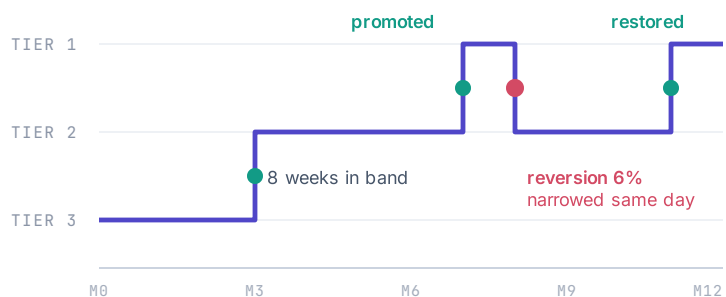
Promotion takes weeks of evidence. Demotion takes one bad fortnight. The asymmetry is deliberate: being slow to widen costs some efficiency; being slow to narrow costs customers.

WHEN APPROVALS STOP BEING EVIDENCE

A 99% approval rate with a two-second median review is a rubber stamp. Microsoft's red team names consent fatigue as a leading way oversight fails.

FIGURE 13 • ONE INTENT, TWELVE MONTHS

Refunds \$100–\$500



Illustrative. The demotion is the system working, not failing. A program that has never demoted an intent is not reading its scorecard.

Autonomy is not a launch setting. It is a monthly decision, made on evidence, in both directions.

Five executives, five questions

Every executive reads the same scorecard for a different question. Give each the two numbers that answer it, and the conversation moves from whether the agent works to what to do next.

FIGURE 14 • THE EXECUTIVE LENS

ROLE	THE QUESTION	TWO NUMBERS	THE DECISION IT OWNS
CEO	Are customers choosing us more because of this?	▪ Net value vs. holdout ▪ Return-to-agent trend	Which journeys go agentic next
CFO	Is the saving real — and will it last?	▪ Cost per resolved intent ▪ Benefits realized vs. committed	Fund the next tranche, or hold
CMO	Is it keeping the brand promise?	▪ Customer effort ▪ Retention, served cohort	What the brand may claim about the agent
CPO	What should the product fix so customers need the agent less?	▪ Top repeat-contact intents ▪ Product-caused contact rate	Roadmap items that remove demand
CX lead	Which intent moves tier next — and which moves back?	▪ Autonomy decisions ▪ Handoff quality	The monthly widen, hold or narrow call

Ten numbers, one source. Every figure comes from the same scorecard and the same definitions, so no two executives argue from different versions of the truth.

The CPO row is the one most often missing

The cheapest resolved intent is the one that never arrives. Repeat-contact data, read by intent, is a product-defect list: every high-volume intent the agent resolves perfectly still asks why customers had to ask.

Three things for the board pack

- **Net value against holdout**, with the attribution method named.
- **The trust trend**, not the trust score.
- **Every autonomy change** this quarter, both directions, with the reason.

One number invites a story. Ten numbers from one definition invite a decision.

Six moves, in this order

None of these needs a new platform, and most run on logs you already keep. Together they make an agent you can defend to a customer, a CFO and a board.

FIGURE 15 • THE SEQUENCE

01 Redefine resolution

Confirmed in the record, no repeat within seven days, any channel.

02 Join the channels

So repeat contact is visible wherever the customer goes next.

03 Set the holdout

5–10% of eligible traffic, before launch or before the next widening.

04 Sign the benefits

Baseline, band, attribution, owner — dis-benefits included.

05 Write the autonomy rules

Expand, hold, reduce, suspend — published before they are needed.

06 Decide monthly

One scorecard, one recorded decision per intent, both directions allowed.

Order matters more than pace. Teams that start at the scorecard build a dashboard for a definition of resolution nobody agreed. Teams that skip the holdout spend the next year arguing about attribution.

What good looks like at day ninety

- One definition of resolution, used by operations and finance alike.
- Repeat contact visible across every channel.
- A holdout running, and an attribution method named.
- A signed benefits case with dis-benefits in it.
- Written autonomy rules, and at least one decision recorded against them.

The caveat that belongs here

Benchmarks for agentic customer service are young and mostly vendor-published. Treat every threshold here — ours included — as a starting band, to be replaced by your own baseline within a quarter.

What does not change is the principle: measure the outcome the customer experienced, against what would have happened without the agent.

Containment tells you the agent was busy. Resolution tells you the customer was served. Only a holdout tells you it was worth it.

THE SERIES

VOLUME I • DESIGN IT RESPONSIBLY

Agentic AI Customer Experience & Innovation

Trust, authority tiers and the patterns that buy autonomy.

VOLUME II • BUILD IT SAFELY

The Agentic CX Build Blueprint & Guardrails

Four phases, nine artifacts and the stop you design first.

VOLUME III • OPERATE IT PROFITABLY

The Agentic CX Performance System

Measure resolution, prove value and earn the next level.

WHERE THIS GOES NEXT

Measure it. Prove it. Earn the next level.

f/cx.ai is a customer experience design and innovation agency for the AI era. We design, prototype and prove the journeys, service models and AI agents enterprise organizations run on.

IF YOU ARE LIVE BUT UNSURE

Conversation Intelligence

We read your agent's conversations and logs against the seven-level hierarchy and show where contained is not resolved.

IF NOBODY OWNS THE NUMBER

Fractional CX Leadership

A senior CX executive, part-time, who owns the scorecard, the benefits case and the monthly autonomy call until your team can.

IF YOU ARE ABOUT TO WIDEN

CX Governance & Readiness

Holdout design, benefits template and autonomy rules — written and signed before the next tier opens.

An agent earns autonomy the way a person does: by being right, visibly, for long enough.
Everything else is activity.

THE CONVERSATION WE HAVE MOST OFTEN

"We launched it — how do we know it is working?" If that is the question in front of your leadership team, it is the one we are set up for.

CONTACT

Skip Vanderburg

Founder & Chief Experience Officer

innovate@fractionalcx.ai

fractionalcx.ai

Where every external number came from

Every figure in this paper is original. External data points are attributed where they appear and listed here in full. Three notes on care: vendor-published performance data is self-interested, so it is paired with independent sources wherever a claim rests on it; several widely quoted loyalty figures are old and are flagged as such; and the Salesforce and Gartner resolution figures on page two measure different things and should not be compared as like for like. Illustrative values in the worked build — volumes, costs, thresholds, cohort results — are constructed and continue the Volume II example.

RESOLUTION, CONTAINMENT AND SELF-SERVICE

Gartner 2024	<i>Survey finds only 14% of customer service issues are fully resolved in self-service. 5,728 customers, December 2023. 36% of "very simple" issues resolved; 45% said the company did not understand their intent; 73% use self-service at some point.</i>
Salesforce 2026	<i>Agentic Enterprise Index. Seven in ten conversations resolved without escalation; +10 NPS on a named concierge deployment. Vendor telemetry.</i>

CUSTOMER EFFORT, LOYALTY AND EXPERIENCE QUALITY

Dixon, Freeman & Toman 2010	<i>Stop Trying to Delight Your Customers. Harvard Business Review. More than 75,000 customers; effort, not delight, drives disloyalty; introduces the Customer Effort Score.</i>
CEB, now Gartner 2013	<i>The Effortless Experience. 96% of customers with high-effort service interactions become more disloyal, against 9% with low-effort ones. Older research; the direction is well replicated, the exact figure less so.</i>
Forrester 2025	<i>Global Customer Experience Index. 469 brands, 13 countries, more than 275,000 customers. 21% of brands declined, 6% improved, 73% unchanged; small CX gains reduce churn and raise share of wallet.</i>
Reichheld, Bain & Company via HBR, 2014	<i>The Value of Keeping the Right Customers. A 5% increase in retention increases profits by 25% to 95%. Old and sector-specific; cited for its shape, not its range.</i>

AI VALUE, SCALING AND ECONOMICS

IBM Institute for Business Value 2025	<i>CEO Study. 2,000 CEOs, 33 countries. 25% of AI initiatives delivered expected ROI over recent years; 16% scaled enterprise-wide.</i>
McKinsey 2026	<i>Rewiring customer experience for the agentic era. Named program outcomes including +24% first-contact resolution; customer-facing deployments scale at 3.5× the rate of other domains.</i>
Salesforce 2026	<i>AI service agents and customer satisfaction. Seven in ten service organizations adopting agents see measurable value within 60 days.</i>
Gartner 2026	<i>GenAI cost per resolution forecast. Cost per resolution to exceed \$3 by 2030, surpassing many offshore human agent costs.</i>

AUTONOMY, OVERSIGHT AND CONTROL

Microsoft AI Red Team 2026	<i>Taxonomy of failure modes in agentic AI systems, v2.</i> Human-in-the-loop bypass among the most consistently exploited failure modes, largely through consent fatigue.
Knight First Amendment Institute 2025	<i>Levels of Autonomy for AI Agents.</i> Autonomy named by the role the person keeps. Basis for the tier language carried from Volume I.
Anthropic 2026	<i>Measuring AI agent autonomy in practice.</i> Autonomy as a property of the deployment, scored against consequence.

Prepared by f/cx.ai. Volume III, following *Agentic AI Customer Experience & Innovation* (Volume I) and *The Agentic CX Build Blueprint & Guardrails* (Volume II). Frameworks here are working instruments and should be revised against your own evidence.