

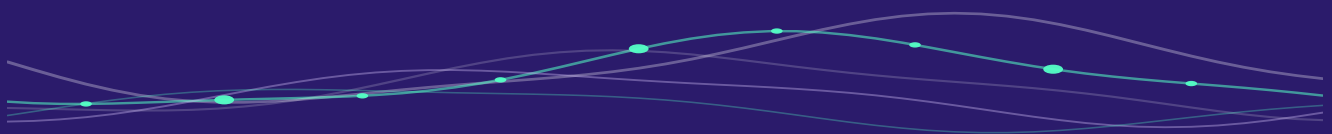
THE ESSENTIAL GUIDE

AI for Customer Service: Self-Learning Agentic CX

How agentic AI actually resolves customer issues, and improves itself

Contents

Executive summary	03
THE MARKET AND WHY NOW	
1. The CX pressure cooker	05
2. The agentic AI inflection	06
3. The reality gap	07
4. Why traditional automation cannot keep up	08
THE SELF-LEARNING AGENTIC CX GUIDE	
5. What agentic and self-learning actually mean	10
6. How it works, step by step	11
7. An illustrative walk-through	16
8. One intelligence, every surface	17
9. Getting cheaper and better over time	18
10. The benefits, quantified	20
11. Enterprise-grade by design	21
12. Why Omilia	23
DECISION TOOLKIT	
Myth versus reality	24
A buyer's checklist	25
Getting started · Frequently asked questions	26
From hype to resolution	27
References	28



Executive summary

The customer experience market has a credibility problem. Gartner® projects that by 2029, agentic AI will autonomously resolve **80% of common customer service issues without human intervention, leading to a 30% reduction in operational costs.**¹ It further states: over 40% of agentic AI projects will be canceled by the end of 2027, due to escalating costs, unclear business value or inadequate risk controls.²

The prize from agentic AI in customer service is real, but so is the risk. Forrester warns that three in 10 firms will harm their total-experience growth with frustrating AI self-service.³ Countless vendors promise that AI will resolve most customer issues almost overnight. However, the difference between real reward and risk is readiness. The winners will not be the organizations that deploy the most AI. They will be the ones that deploy AI that learns from real customer interactions, continuously improves itself, and retains human control.

This paper covers, in depth, three things:

- 1** Where the market is heading and what the numbers really say
- 2** Why traditional automation cannot keep up
- 3** How Omilia's self-learning agentic CX works, what it delivers, and how it changes the economics of your entire contact center.

This guide provides value to both operational and technical teams. CX and contact center leaders will find the outcomes they care about, containment, CSAT, handle time, cost-to-serve, and time-to-value, while IT and technology leaders will discover detailed insights on safety, model governance, security, integration, and graceful escalation. Every business benefit is paired with a technical note, ensuring the content is relevant for every stakeholder.

The pages that follow explain which CX AI genuinely learns, acts, and stays governable, and how to tell it apart from what does not.

The winners will not be the organizations that deploy the most AI. They will be the ones that deploy AI that learns from real customer interactions, continuously improves itself, and retains human control.

The market and why now

Where the market is heading and what the numbers really say, and why traditional automation cannot keep up.

1. The CX pressure cooker

Customer expectations have quietly become unforgiving. People now compare every service interaction not against your competitors, but against the best digital experience they had all week, from any company in any industry. They expect fast answers, no repeated questions, and resolution on the first contact. When your contact center falls short, they do not complain. They leave.

The numbers are stark. PwC's 2025 Customer Experience Survey, titled "The Loyalty Illusion," found that 52% of consumers stopped buying from a brand after a bad experience, and that 86% still consider human interaction essential to their brand experience.⁵ Zendesk puts a finer point on it: 73% of customers will switch to a competitor after multiple bad experiences, and more than half will leave after just one.⁶

There is also a dangerous blind spot. PwC found that around nine in ten executives believe customer loyalty has grown in recent years, while only four in ten consumers agree.⁵ Leaders think they are winning while customers quietly churn.

Leaders think they are winning while customers quietly churn.

At the same time, economics cuts both ways. By 2030, cost per resolution for generative AI (GenAI) will exceed \$3, higher than many B2C offshore human agents, according to Gartner, Inc.⁷ Rising data center costs, a pivot from subsidized growth to profitability for AI vendors, and increasingly complex use cases that consume more tokens and require expensive talent, will lead to soaring AI costs for customer service organizations. Automating naively is not a cost story; automating well is.

And the money is following: IDC predicts that the worldwide conversational artificial intelligence (AI) software services market will grow to \$44.7 billion in 2030 at a total market compound annual growth rate (CAGR) of 38.2%.⁸

This is a structural shift in how service gets delivered, and it is happening whether any single organization is ready or not. Automation is no longer optional. The question is not whether to automate, but how to automate without becoming one of the brands customers flee.

2. The agentic AI inflection

For a decade, contact center AI could talk, though it had a very hard time trying to listen. A chatbot could answer a question and, at best, hand the customer a phone number or a link. Now, agentic AI can effortlessly converse with humans in an engaging manner. It can reason about what the customer needs, plan the steps to resolve that, and take real action across backend systems: for example, verifying an identity, looking up an account, processing a payment, updating a record, issuing a refund.

Gartner predicts that by 2028, 70% of customer service journeys will begin – and be resolved – in conversational, third-party assistants built into their mobile devices.⁹ The adoption curve is steep in the near term too: Gartner expects 40% of enterprise applications to feature task-specific AI agents by the end of 2026, up from less than 5% in 2025.¹⁰

The difference between answering and resolving is the difference between deflection and genuine service.

3. The reality gap

Part of the problem is noise. Gartner calls out “agent washing,” where vendors rebrand old chatbots, scripts, and robotic process automation as agents, and estimates that only around 130 of the thousands of agentic AI vendors are the real thing.⁴

This is why independent recognition matters, and why Omilia’s placement as a [Visionary in the 2026 Gartner® Magic Quadrant™ for Conversational AI Platforms](#), and a [Leader in The Forrester Wave™: Conversational AI Platforms For Customer Service, Q2 2026](#), is more than a badge. It separates a proven, full-stack platform from the rebadged crowd.

So the paradox is clear. The market rewards moving fast on agentic AI, and it punishes moving carelessly. A failed self-service launch does not just fail to help. It actively erodes the trust and revenue you were trying to protect. The organizations that win will be the ones that can be both fast and safe. That requires a fundamentally different way of building automation.

The market rewards moving fast on agentic AI, and it punishes moving carelessly.

4. Why traditional automation cannot keep up

Building a contact center bot the traditional way is a craft. A team of designers, linguists, and engineers reads through transcripts, maps out conversation flows, codes decision trees, and launches them. Every use case, whether a payment, a balance check, or a card replacement, gets its own design cycle. When products change or policies shift, the whole cycle repeats. That model has three built-in weaknesses.

- **It does not scale.** The effort to build and maintain automation grows in step with the number of use cases. More coverage means more people, not fewer.
- **It is slow.** Weeks or months can pass between spotting a pattern and shipping automation for it. By the time it launches, the behavior it captured may have already moved on.
- **It relies on guesswork.** It depends on humans interpreting what a typical call looks like. That introduces bias and misses the statistical detail that only shows up across hundreds of real interactions.

Pure large language model agents solve the speed-to-build problem but introduce a new one: the cost of tokens. LLMs are expensive to run on every step of every call, and left unmanaged they drift, which is a major reason so many projects stall. The market has effectively been forced to choose between cheap-but-rigid scripted automation and flexible-but-costly autonomous agents. Omilia's Self-learning CX Agents refuse that trade-off, and the rest of this paper explains how.

The self-learning agentic CX guide

What agentic and self-learning actually mean, how the loop works step by step, and what it delivers.

5. What agentic and self-learning actually mean

Two words get used loosely, so it is worth being precise. Agentic means the system can take action on the customer's behalf, not just answer questions. Self-learning means the system builds and improves itself from real interactions, rather than depending on people to hand-design and hand-tune every flow. Put them together and you get automation that resolves issues and gets better on its own.

It helps to see where this sits on the path contact centers have traveled.

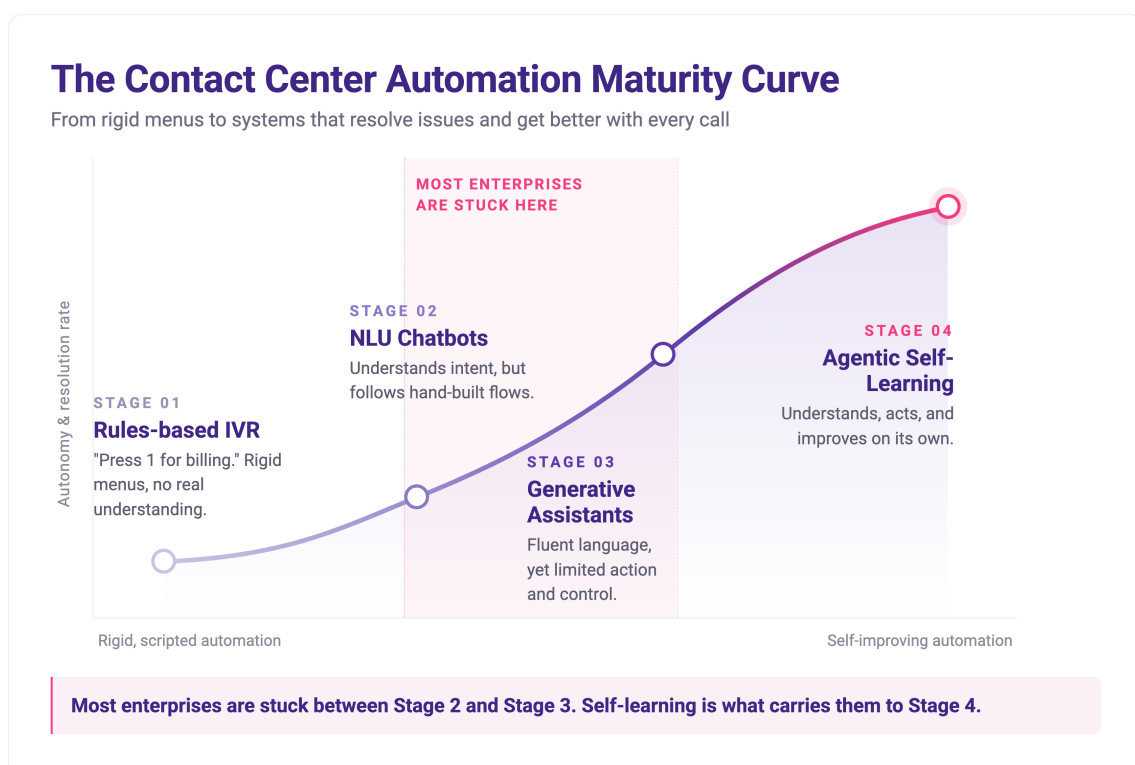


Figure 1. Most enterprises are stuck between Stage 2 and Stage 3. Self-learning is the bridge to Stage 4.

The first generation was rules-based IVR: rigid press-one-for-billing menus that understood nothing and forced customers to adapt to the machine. The second was NLU chatbots that could recognize intent but still followed the flows a human had to build and maintain by hand, so coverage stayed narrow and maintenance never ended. The third was generative assistants that spoke fluently but often lacked the ability to take action, or the governance an enterprise requires, which is how eloquent bots still failed to resolve anything. The fourth generation, agentic self-learning, understands, acts, and improves autonomously, with humans supervising rather than hand-building.

The reason so many organizations are stuck between the second and third generations is that each earlier stage puts humans on the critical path for building and maintaining automation. Self-learning removes that bottleneck by learning the work directly from the interactions you already have. Think of it less like buying a robot and more like hiring a fast learner who studies your best people, drafts a plan for how to handle each type of call, and asks for your sign-off before taking action.

6. How it works, step by step

There are two things worth understanding: how the platform handles a single live conversation, and how it builds and continuously improves the agents behind those conversations. The second is the self-learning loop, and it is where the real difference lies.

Handling a live conversation

Every interaction starts with understanding what the customer actually wants, across accents, phrasing, and noisy phone lines. This runs on Omilia's proprietary speech recognition and language understanding engine. Because Omilia builds its own speech models rather than relying on third-party technologies, intent understanding stays tuned to your domain and consistent across channels. The agent then reasons about how to resolve the request, and follows the successful path the platform has learned from how your best human agents handle that type of call. To complete the task, it acts across your systems, calling the APIs needed to look up an account or process a payment, connecting natively to your systems and to CRMs. Predictable steps are run by lightweight, deterministic agents, with answers provided by the platform's knowledge base. When a request falls outside the agent's scope or identifies a sensitive task, it escalates gracefully to a human with full context. Throughout this process, the platform's Co-Pilot guides the people building and supervising these agents.

The self-learning loop

Behind those live conversations runs a continuous loop of five stages. Each turn makes the agents a little smarter, and a human stays in control at every point that matters.



Figure 2. The self-learning loop. AI drives each stage; humans approve at every checkpoint that matters.

Stage 1: Monitor

The loop begins with your data. All audio or text transcripts are ingested and processed, and you do not need to label, clean, or organize anything first. The platform transcribes the audio, separates the agent from the customer, and reads each conversation to extract structured detail: what the customer wanted, whether it was resolved, the sentiment, and the ordered sequence of steps the agent took. Ingestion is deliberately simple, and results start appearing within minutes. As calls are processed, use cases such as FAQs, account management, and billing emerge on their own, with no one telling the system those categories exist.

Input	Type	Notes
MP3, WAV, M4A	Audio	Speaker separation is automatic
TXT	Transcript	Uses Customer and Agent speaker labels
JSON	Pre-transcribed	Conforms to the OCP transcript schema
PDF, web URL	Knowledge	Enriches the answer knowledge base
OpenAPI spec	API definition	Suggests and connects backend tools

What turns a free-form call into something a machine can learn from is structure. As it reads each conversation, the platform classifies every move the agent made into a small set of typed steps: gathering information from the customer, confirming details, announcing information back, and performing a backend transaction. A payment call, for example, becomes an ordered sequence such as gathering account number, confirming identity, announcing balance, and transacting payment. That typed sequence is the raw material for everything that follows.

Stage 2: Analyze

A single call is an episode. Hundreds of calls are a pattern. The platform clusters conversations by topic and lines them up to find the steps that appear in almost every successful call of a given type, and the branches that only appear in specific situations. The technique is borrowed from genetics. Just as biologists align DNA sequences to find the regions conserved across species, the platform aligns conversation sequences to find the steps conserved across calls. When most “make a payment” calls follow one path but a quarter divert to collections because the account is overdue, the system recognizes both as valid routes and records the condition that triggers each.

It waits until it has seen enough variation before proposing anything: a minimum of five conversations per use case to begin, and a few hundred to a thousand for production-quality results. Building an agent from a single call would be like training a new hire by having them listen to one phone call. The output of this stage is a resolution strategy for each use case, a structured, human-readable description of how the call should flow, what information to collect and in what order, which steps need explicit confirmation, which backend systems to call, and the safety rules and business policies your agents consistently follow.

You see the business case before you commit

For every use case, the platform forecasts projected AI resolution rate, CSAT, handle-time reduction, and cost savings before you deploy anything. Decisions are made on evidence, not on a leap of faith.

Predicting the next best action

Once the platform has learned enough patterns, it can predict what should happen next in a live conversation, a little like autocomplete for customer service. Given the conversation so far, it estimates the most likely next step and attaches a confidence score. Collecting the date of birth after the account number might carry very high confidence because it followed that order in the vast majority of similar past calls. Those predictions do useful work in three places: they appear as real-time suggestions to human agents, they bias AI agents toward proven paths while still allowing sensible deviation, and at the highest confidence levels, they let the system hand a step to a fast, deterministic module with no AI cost at all. This is also the mechanism that makes convergence, described in Section 9, possible.

Analyze doesn’t average every agent’s calls equally. When lining up conversations for a use case, it weights the paths taken by agents whose calls are resolved cleanly and leave customers satisfied, so the resolution strategy it drafts reflects how the center’s strongest performers actually handle the interaction. From there, that strategy does double duty: an AI agent can execute it directly to resolve the interaction end to end, or it can surface as a next-best-action to a newer agent still building judgment, the same guidance a senior colleague would give in the moment. Every call that resolves well, whether handled by AI or a coached human, feeds back into the pattern the platform learns from, so the standard it teaches rises with each interaction rather than staying fixed.

Stage 3: Build and Optimize

Once a resolution strategy is approved, the platform generates the complete agent automatically, with no prompt engineering on your side. From what it learned, it produces the agent's identity and scope, the safety guardrails such as never process an account change without authentication, the business policies with their priorities and triggers, the step-by-step conversation flow, the backend tools to call, the factual knowledge to draw on, and even the phrasing, ranked by which wording led to the best outcomes in real calls. Your agent ends up speaking the way your best human agents speak.

Three kinds of agents work together. A Concierge agent is the front door: it greets every caller, answers common questions instantly, and routes complex requests to the right specialist, updating itself automatically as you add specialists. FAQ agents handle fast, accurate questions and answers, powered by the knowledge base that traces back to a real source rather than being invented. Task agents run self-service flows such as billing, account management, and transactions, following the learned conversation flow and calling your backend APIs to actually complete the work.

The control plane: who handles what

Not every moment in a conversation needs the same intelligence. Some steps are highly predictable; others require genuine judgment. A control plane makes that routing decision in real time. When confidence is high, a lightweight deterministic module handles the step with maximum efficiency and no AI cost. When confidence is low, the full AI agent takes over. And for any sensitive action with real-world consequences, a human-in-the-loop safeguard requires approval before execution. The customer never notices these handoffs. They simply get a smooth, reliable experience while the platform quietly optimizes for cost and reliability underneath.

Stage 4: Evaluate

Nothing reaches production untested. Before deployment the platform generates a set of test scenarios and runs them with one click. These are not keyword checks. It creates synthetic customers, each with a persona and a goal, who hold multi-turn conversations with your agent, and it scores goal completion, sub-task accuracy, and conversation quality. The scenarios cover the happy path, deliberate guardrail-violation attempts, escalations that should route to a human, unusual-but-valid edge cases, and adversarial stress tests designed to break the agent. All tests must pass before deployment is confirmed, and you review the projected KPI impact side by side one last time before deciding.

Stage 5: Launch

Deployment is a click. The platform validates the configuration, provisions the application, wires up the agents and routing, connects the tools, and sets the status to active. For a standard use case with enough data, the journey from upload to a live, production-grade agent takes under 15 minutes, not weeks or months, and a pause control gives you an instant off-switch at any time. Then the loop closes: every interaction the live agent handles feeds straight back into Monitor, and the cycle begins again with more evidence than before. Structural changes, such as a new phase appearing in a use case, always surface for human review, and every version is tracked so you can see exactly what changed and step back if needed.

Where humans stay in control

Automation without oversight is exactly what makes so many AI projects fail, so self-learning is built the other way around. Nothing is deployed without your approval. You accept or reject every extracted answer, review every resolution strategy, and sign off on every new version before it affects live traffic. For sensitive actions with real-world consequences, such as processing a payment or changing an account, a human-in-the-loop safeguard requires approval before execution, regardless of whether an automated step or an AI agent initiated it. The customer never sees these handoffs. They just get a smooth, reliable experience.

See it on your own data

The fastest way to judge self-learning is to watch it learn from your calls. [**Book a demo or talk to an Omilia expert**](#) to see the use cases the platform finds and the impact it forecasts, before you commit a single resource.

7. An illustrative walk-through

To make the loop concrete, here is an illustrative example. The scenario is representative rather than a specific customer, and it mirrors the kind of use case the platform routinely discovers.

A utility connects a few dozen recorded billing calls. Within minutes, **Monitor** has transcribed them, separated agent from customer, and surfaced three use cases it was never told to look for: general FAQs, account management, and billing inquiries. Reviewing the FAQ candidates, a supervisor sees a clean question and answer about the utility's EV charging program, extracted directly from a real call, and clicks accept. That answer is now in the knowledge base, traceable to its source.

As more calls arrive, **Analyze** has enough billing conversations to line them up and find the dominant path: authenticate the caller, verify the service address, retrieve the balance, then either take a payment or, when the account is overdue, move to a collections branch. It drafts a resolution strategy capturing that flow, the confirmation steps, the backend lookups, and a policy the agents never break, which is that account changes require authentication first. Alongside it, a KPI forecast projects the resolution rate, CSAT, handle-time reduction, and monthly savings for the use case.

Evaluate then spins up synthetic callers, one paying on time, one past due, one trying to skip authentication, one asking something unusual, and confirms the agent handles each correctly. A supervisor approves, and with one click a Task agent goes live in minutes, with a Concierge updating itself to route billing questions to the new specialist.

Nothing here required a designer to script a dialog, an engineer to write a prompt, or a data scientist to label a dataset. The knowledge was already in the calls. Over the following weeks, as thousands more billing calls flow through, the strategy versions up on its own, the predictable steps shift to deterministic execution, and the cost per interaction falls while the resolution rate climbs.

8. One intelligence, every surface

Everything the loop learns is stored once, as a resolution strategy: the steps, the order, the confirmations, the backend calls, the policies and guardrails. It describes how a use case should be handled in a form that is independent of who or what carries it out. That independence is the platform's most consequential design decision, because it means one learned intelligence can serve every modality and power every contact center capability without being rebuilt for each. Customers get identical treatment wherever they arrive, and you maintain one thing, not four.

The same strategy then drives three capabilities. It runs an autonomous AI agent that resolves the interaction end to end. It powers Agent Assist, surfacing the next best action with a confidence score to a human handling the same call. And it drives automated quality management, because a description of a correct interaction is a scoring rubric as much as a build specification. Traditional QM samples one or two percent of interactions against a form somebody wrote; here every interaction is scored against the learned strategy, so human and AI agents are held to the same standard, so you can compare like with like.

One learned intelligence can serve every modality and power every contact center capability without being rebuilt for each.

Note where that guidance comes from. The strategy is learned from the interactions that were actually resolved, which in practice means it is learned from your most experienced agents. That expertise then reaches a new hire on their first shift as a real-time suggestion, without a senior agent writing documentation, sitting in a coaching session, or lifting a finger. The gap between your best people and your newest closes on its own, and every turn of the loop raises the floor for the entire contact center.

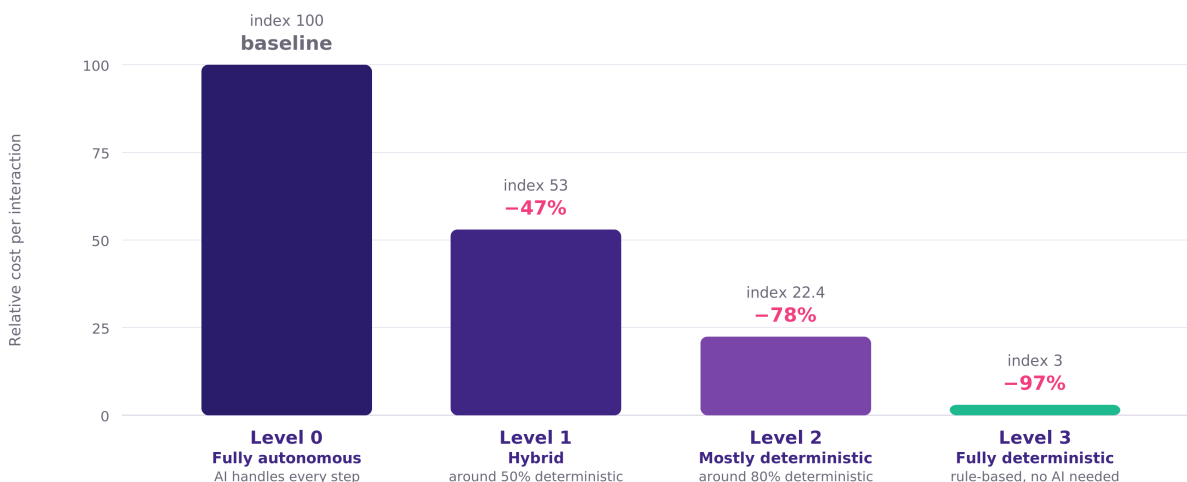
Each surface feeds the same loop, so improving one improves all three. Bought separately, as most organizations do today, they are three data models, three configuration efforts, and no shared learning. Here, once a strategy is approved, standing up an AI agent, enabling Assist for the teams handling that use case, and switching on automated QM across both are configuration choices, not projects, one click per capability.

9. Getting cheaper and better over time

Most technology gets more expensive to run as you use it more. Self-learning does the opposite, and this is the part that changes the business case entirely. Autonomous AI agents give excellent conversational quality but cost the most to run, because intelligence is applied to every step of every call. Deterministic automation, the scripted and rule-based kind, runs one to two orders of magnitude cheaper, but traditionally took heavy upfront engineering to build. Self-learning resolves that tension in an elegant way: the autonomous agents generate exactly the data needed to build their own cheaper, deterministic replacements.

Cost per interaction falls as use cases converge

Relative cost per interaction, indexed to fully autonomous handling = 100



Reaching Level 2, around 80% deterministic, cuts handling costs by up to 78%.
Full convergence to Level 3 reaches a 97% reduction, at the same customer experience.

Figure 3. Cost per interaction falls as a use case converges from Level 0 to Level 3. Indexed to fully autonomous handling = 100.

As the platform sees more calls for a given use case, it progressively hands the predictable steps to fast, low-cost deterministic execution and reserves the expensive AI for the genuinely hard moments. This graduated shift is called convergence.

Level	How the use case is handled	Relative cost
Level 0	Fully autonomous. AI handles every step.	Highest
Level 1	Hybrid. Deterministic answers, AI for complex steps.	Reduced
Level 2	Mostly deterministic. AI reserved for edge cases.	Low
Level 3	Fully deterministic. Rule-based, no AI needed.	Lowest

For a deployment handling five million interactions a month, moving to a hybrid steady state of roughly 80% deterministic can cut handling costs by **up to 78%**, and full convergence pushes that toward **97%**. Quality rises at the same time. Interactions that flow smoothly confirm the pattern. Interactions that fail to route to the AI agent become new observations, and are folded back in as improvements. The agent you deploy today quietly becomes faster, cheaper, and more accurate tomorrow, with no extra engineering, no manual intervention, and no visible change for the customer.

10. The benefits, quantified

The advantages land differently depending on where you sit, so here is a view for both sides of the buying committee. Every one of them traces back to the same source, the learning loop, which is why the gains compound rather than plateau.

For CX and operations leaders

- **Speed to value.** New use cases go from idea to a live, tested agent in minutes rather than months, so you respond to demand instead of waiting on a backlog.
- **Higher containment, better CSAT.** Agents are built from the paths that already resolve issues successfully and speak the way your best people do, so they resolve more on first contact, and graceful escalation means customers are never trapped.
- **Lower cost to serve.** Convergence bends the cost-per-interaction curve downward over time rather than up.
- **Forecast before you fund.** KPI projections let you make the business case with numbers before committing resources.
- **Consistency at scale, in your voice.** Every agent follows the patterns and policies your best people already use, on every call, in every language, with wording drawn from your own successful calls.

For technology teams

- **No manual flow design or prompt engineering.** The platform builds the configuration from your data and maintains it as data changes.
- **Automatic, rigorous testing.** Synthetic personas stress every agent across happy paths, guardrails, escalations, edge cases, and adversarial attacks before launch.
- **Real integrations.** Native RingCX support, prebuilt CRM connectors for Salesforce, Microsoft Dynamics, Zendesk and more, plus open APIs and SDKs. Backend tools are suggested automatically from what your agents actually do.
- **Versioned and reversible.** Every change is tracked and surfaced for review, so you always know what changed and can step back.
- **Continuous, granular improvement.** Failed or escalated interactions are analyzed automatically at every layer of the stack, so problems get diagnosed precisely rather than guessed at.

11. Enterprise-grade by design

For IT, digital, and CIO buyers, capability is only half the decision. The other half is whether the platform can be trusted inside a regulated enterprise, and whether it will integrate without becoming another stalled project. Self-learning treats governance as a first-class feature, not an afterthought.

Security and compliance are foundational. Omilia holds and supports the certifications regulated industries require, including **PCI-DSS, SOC 2 Type II, ISO 27001, HIPAA, GDPR, and CCPA**, and supports FedRAMP-aligned deployments. Forrester specifically noted Omilia as a [good fit for highly regulated, security-first organizations](#), citing its centralized model governance and strong AI guardrails.

Governance is layered. A human reviews and approves what an agent will do before it ever handles a live customer. Guardrails and business policies are learned from how your own agents consistently behave, so the rules reflect your real operations rather than a generic template. Every extracted answer traces back to a real conversation or an approved document, so it is grounded rather than invented. Sensitive actions require explicit human sign-off regardless of how the step was initiated, and every proposed change to a live agent is versioned and surfaced for review before it takes effect.

Self-learning treats governance as a first-class feature, not an afterthought.

The advantage of owning the whole stack

When a call goes wrong, most vendors can only guess why, because they stitch together third-party speech and language services they cannot see inside. Omilia owns the entire conversational AI technology stack, from speech recognition and language understanding through dialog to speech synthesis. That means when an interaction fails or is escalated, the platform can pinpoint the cause across three layers.

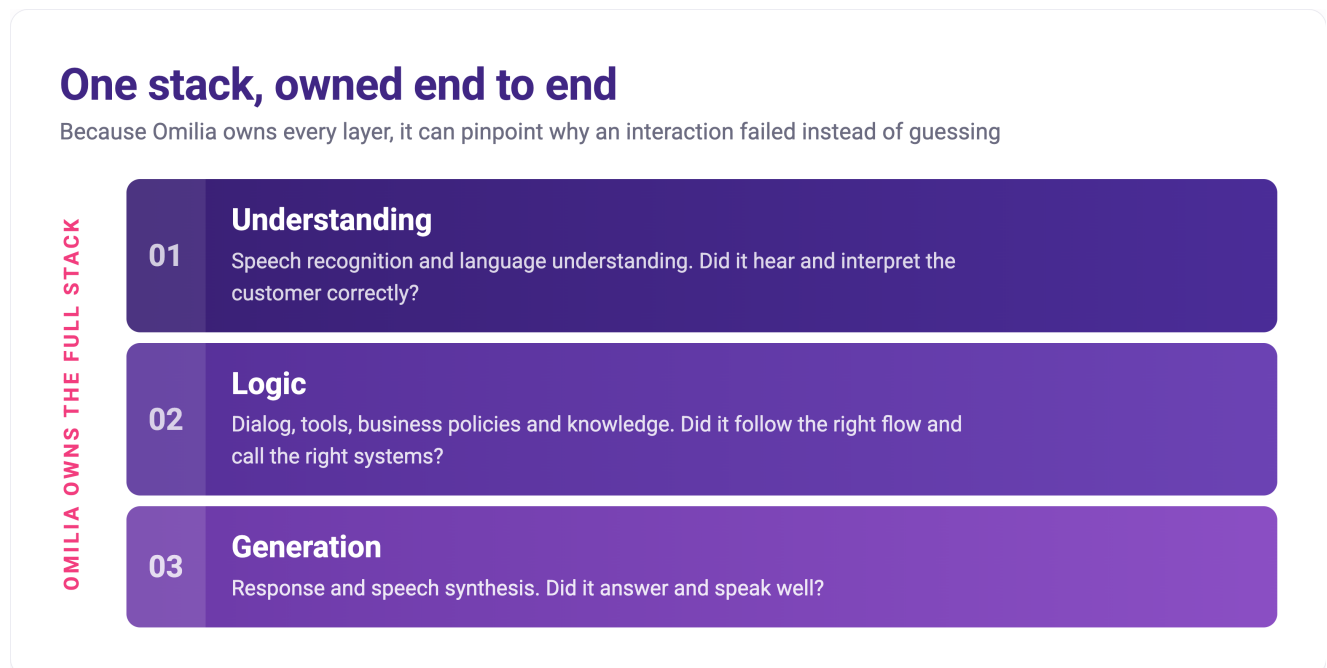


Figure 4. Full-stack ownership lets the platform diagnose failures precisely rather than guess.

Precise diagnosis is why Self-Learning CX Agents keep getting better instead of plateauing. A problem in understanding, logic, or generation is identified at its layer and fixed at its root, rather than papered over with a broad prompt change that risks new failures elsewhere. It is also what makes convergence safe: the platform can only crystallize a step into cheap deterministic execution because it can see, precisely, that the step behaves reliably.

12. Why Omilia

Plenty of vendors can give you an autonomous agent. Very few can give you one that also gets cheaper and more reliable the longer it runs. That difference comes down to owning the full stack: speech, language understanding, dialog management, the deterministic runtime, and the self-learning layer that ties them together.

A pure-play AI vendor can build an autonomous agent but has no deterministic runtime to convert it into cheaper execution, so it stays stuck at the most expensive level. A traditional contact-center vendor owns the deterministic runtime but cannot get from observation to a working autonomous agent without months of manual engineering. Omilia does both, and moves smoothly between them. That end-to-end ownership is what turns a one-time automation project into a system that compounds, and it is a large part of why Omilia believes Forrester named them a Leader in its 2026 evaluation, citing proprietary speech models, tightly governed agents, and support for standards up to FedRAMP.

That end-to-end ownership is what turns a one-time automation project into a system that compounds.

DECISION TOOLKIT

Myth versus reality

Myth: “AI will resolve 80% of our issues the day we switch it on.”

Reality: Gartner's 80% is a 2029 projection for common issues,¹ and Forrester warns that 2026 rewards preparation. Self-learning closes the gap by starting from patterns already proven in your contact center.

Myth: “Agentic AI means handing control to a black box.”

Reality: Humans approve every agent before launch, sensitive actions require sign-off, and every new version is surfaced for review.

Myth: “More AI everywhere is always better.”

Reality: The win is AI that converges to cheap, deterministic execution where a step is proven. AI on every step just means higher cost with no added value.

Myth: “All agentic vendors are basically the same.”

Reality: Plenty of vendors can give you an autonomous agent. Very few can give you one that also gets cheaper and more reliable the longer it runs.

A buyer's checklist: questions to ask any agentic CX vendor

Use these to cut through the noise, including the agent washing Gartner warns about. They surface the difference between a system that merely talks and one that learns, acts, and stays governable.

- 1 Does it learn from our own interactions, or must every flow be hand-built?** If the answer is manual design, you have bought a maintenance burden, not a solution.
- 2 Who owns the stack?** Can the vendor diagnose failures across speech, understanding, logic, and generation, or does it depend on third-party parts it cannot see into?
- 3 Can a human approve what the AI will do before it goes live?** And does a person stay in the loop for sensitive actions? Governance should be built in, not promised later.
- 4 How does it get cheaper over time, not just at launch?** Look for a clear path from expensive AI handling to cheap deterministic execution as steps are proven.
- 5 What is the real time-to-value?** Measure from your data to a live, tested agent. Weeks and months are a red flag when minutes are possible.
- 6 Does it meet our compliance bar?** Confirm PCI-DSS, SOC 2 Type II, ISO 27001, HIPAA, GDPR, and CCPA, plus centralized model governance.
- 7 How does it fail gracefully?** Ask how escalation to a human is handled, and whether escalation and adversarial cases are tested before every deployment.

Getting started

You do not need a data science team or a six-month program to begin. The fastest path has four moves: bring your history by pointing the platform at recordings or transcripts you already have; stand up a Concierge as soon as answers are extracted; add task agents use case by use case as the platform gathers enough calls for each; and let the loop run, reviewing suggested versions while resolution rates climb and cost per interaction falls.

A useful rule of thumb for data volume: five conversations per use case is the floor to start learning, a few hundred is where automation becomes reliable, and around a thousand per use case is the target for production-quality results. Self-Learning CX Agents can also run alongside an existing Omilia implementation, so this does not have to be a rip-and-replace program.

Frequently asked questions

How is this different from a normal conversational AI agent? Traditional rule-based agents are a subset of the wider self-learning framework. Self-Learning CX Agents improve continuously from interaction data, which removes the need for manual flow design and ongoing prompt engineering.

Are the agents fully autonomous? They can be, depending on the use case and how you configure them. You can also choose to fall back to a fully autonomous agent instead of escalating to a human.

Does it work for our industry and size? Yes. The framework scales from small teams to large multinationals, with prepackaged starting points for retail, financial services, healthcare, technology, utilities, and travel.

Can we add it to our existing Omilia setup? Yes. Self-Learning CX Agents can run alongside your current implementation, and your Omilia account team can advise on the specifics.

Who decides what goes live? You do. Every suggestion is reviewed and approved by a human before it reaches customers, and sensitive actions require explicit approval to execute.

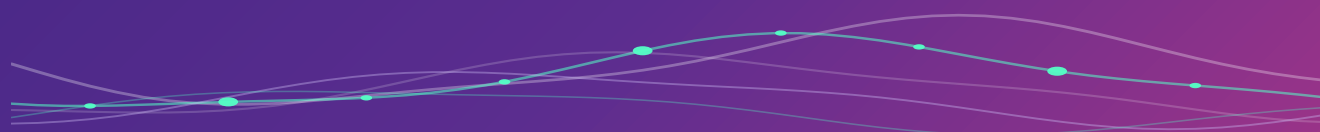
From hype to resolution

The market is loud, and the temptation is to either believe the 80% promise unquestioningly or dismiss it as noise. The honest answer sits in between. Agentic AI genuinely will resolve the majority of common issues, and moving too fast without readiness genuinely will cost you customers and stalled projects. The bridge between the hype and your contact center's reality is not more AI. It is AI that learns from what already works, improves itself continuously, and never acts without human approval.

That is the promise of Omilia's self-learning agentic CX. Autonomous where it is proven, self-improving by design, and enterprise-safe throughout. Your calls already contain everything your AI agent needs to know. Self-learning simply unlocks it.

Talk to an expert

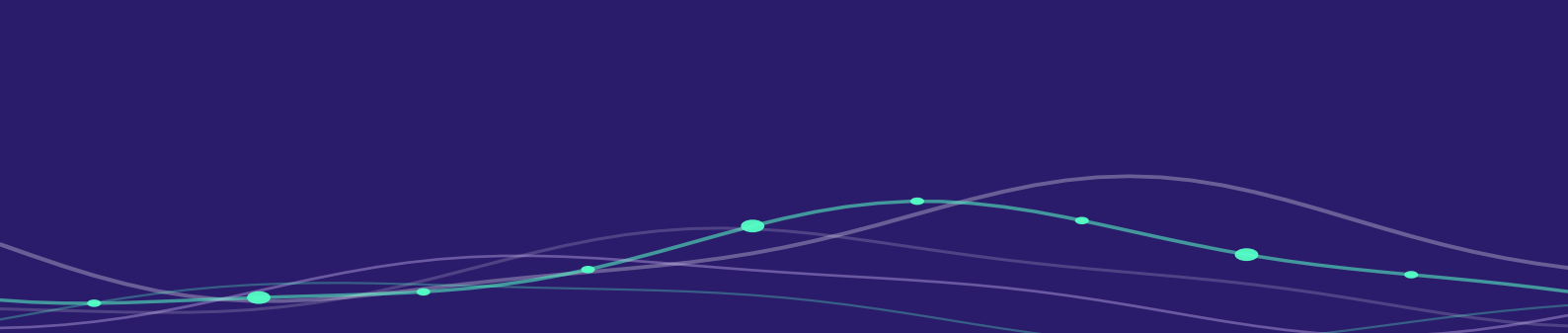
Book a demo or talk to an Omilia expert at omilia.com.



References

1. Gartner press release, Gartner Predicts Agentic AI Will Autonomously Resolve 80% of Common Customer Service Issues Without Human Intervention by 2029, 5 March 2025.
<https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-predicts-agentic-ai-will-autonomously-resolve-80-percent-of-common-customer-service-issues-without-human-intervention-by-20290>
2. Gartner press release, Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, 25 June 2025.
<https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
3. Forrester, Predictions 2026: CX Teams Look To Escape The Orbit Of Dysfunction, 28 October 2025.
<https://www.forrester.com/blogs/predictions-2026-cx-teams-look-to-escape-the-orbit-of-dysfunction/>
4. Gartner. Emerging Tech: AI Vendor Race: Avoid Agentic AI Failure: Build Success Using Right Use Cases, 15 May 2025, ID G00825224, by Anushree Verma.
<https://www.gartner.com/document-reader/document/6478739?ref=askgartner-smart>
5. PwC, The Loyalty Illusion: PwC 2025 Customer Experience Survey.
<https://www.pwc.com/us/en/services/consulting/commercial-excellence/library/2025-customer-experience-survey.html>
6. Zendesk, Customer Service Statistics / CX Trends Report 2026.
<https://www.zendesk.co.uk/blog/customer-service/satisfaction/customer-service-statistics/>
7. Gartner press release, Gartner Predicts GenAI Cost Per Resolution for Customer Service Will Exceed Offshore Human Agent Costs by 2030, 26 January 2026.
<https://www.gartner.com/en/newsroom/press-releases/2026-01-26-gartner-predicts-genai-cost-per-resolution-for-customer-service-will-exceed-offshore-human-agent-costs-by-2030>
8. IDC, Market Forecast: Worldwide Conversational AI Software Forecast, 2026–2030 (Doc #US54086426, August 2026).
<https://my.idc.com/getdoc.jsp?containerId=US54086426&pageType=PRINTFRIENDLY>
9. Gartner press release, Traditional Customer Service Channels Are Losing Ground to Mobile and AI Innovations, 10 February 2025.
<https://www.gartner.com/en/newsroom/press-releases/2025-02-10-traditional-customer-service-channels-are-losing-ground-to-mobile-and-ai-innovations>
10. Gartner press release, Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025, 26 August 2025.
<https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025>

GARTNER is a trademark of Gartner, Inc. and/or its affiliates.



Omilia

Conversational Intelligence

Omilia helps the world's largest enterprises transform customer service with AI. Its native Self-Learning CX Agents automate interactions with precision, empower human agents in real time, and deliver seamless, personalized experiences across all channels – learning from the entire customer journey to unlock continuous improvement beyond the containment “glass ceiling” of legacy siloed models. Built on over two decades of proprietary Agentic AI innovation, with multi-layered anti-fraud capabilities built in, Omilia is trusted by enterprises including Capital One, Discover, Taco Bell, RBC, DWP, First Financial Bank, Purolator, and PSEG to deliver measurable outcomes: lower costs, higher efficiency, and unmatched customer satisfaction, all while preserving the human touch where it matters most.

© 2026 Omilia Natural Language Solutions Ltd.