Spotlight, sponsored by:
Conviva

The AI Agent Performance "Blind Spot": Customer Experience

WRITTEN BY:
Tim Law
Research Director, AI and Automation

PUBLISHED:
November 2026

Explore why AI agents that pass every evaluation can still fail the customer, and what enterprises must measure instead.

I. Introduction

"Without a complete picture, when it comes to customer experience, the current state of agent performance evaluations is at best inferred, and at worst a guess."

As we learn more about AI agent behavior in real-world interactions with customers, enterprises are beginning to realize that a significant gap often exists between how we measure technical performance and how we measure customer experience (CX). This gap defines the difference between what we can observe about agents versus what customers experience. Crucially, that experience is not captured by surveys or thumbs-up ratings. It is revealed by what customers do during and after the interaction, much of which never appears in the transcript.

Measuring the quality of real-life user experience may be the most direct signal for continuously improving agents. Most evaluation and observability platforms focus on such aspects as tool-calling accuracy, task completion, latency, cost, and other technical performance measures that are important for the business to know, but do not provide direct, observable impacts on the customer experience. Without the right analytical framework, it is difficult to know whether the agent fulfilled the customer expectations. What is lacking is any direct measure of the customer's actual experience during the interaction.

Enterprises are increasingly aware of this "blind spot" in their measurement of AI agent performance and need to implement platforms that can provide deep insights into both technical performance and customer experience delivered by agents.

At a glance

Key stats

The key goals for deploying agents in customer service organizations are to achieve three primary objectives:

- Augment the human workforce with AI agents (57%)
- Move beyond manual tasks (56%)
- Improve competitive advantage (54%)

Key takeaway

The focus on efficiencies as the primary objective is a symptom of a "technical" focus on AI agent benefits versus a "customer centric" focus. Enterprises need to pursue objectives such as improving customer experience, loyalty, and retention when measuring AI agent performance. And they need real-time experience intelligence to do it.

The agent evaluation fallacy

Enterprises often mistake AI agent performance on accuracy and latency as satisfying customers' requirements. This is a faulty assumption. Speed and accuracy measure the "correctness" of a transaction but limit enterprises' visibility into the impact on efficiency, customer satisfaction, and customer experience.

Inferring that customers had a great experience with your AI agents based on current technical metrics like speed, accuracy, feedback, and resolution is essentially guessing. It fails to capture the actual customer experience and puts customer relationships at risk. Thumbs-up responses can be rare and may skew negative, and "resolved" can include abandoned chats. Resolved is not the same as "it worked."

Measuring customer satisfaction in an AI agent transaction requires measuring every interaction in production, then slicing by any dimension to find where experience breaks down. Averages hide the segments that fail. Averages fail to capture "agent experience," meaning whether the customer got what they came for and whether they are satisfied. Measuring satisfaction also requires baseline key performance indicators (KPIs) based on fine-grained customer cohorts.

This level of analysis gives enterprises the ability to measure whether interactions are not just "correct" but "efficient and effective." Gaining that understanding requires specific metrics such as:

- **Turns to understanding intent:** How many exchanges it takes before it's confirmed with the user that the agent understands what they came for
- **Clarification rate:** How often the user must stop and restate what they meant
- **Correction rate:** How often the user corrects the agent or redirects it mid-task
- **Scope-narrowing rate:** How often the user shrinks the request to something the agent can handle
- **Downstream rework rate:** How often the user redoes the agent's work afterward, often somewhere else
- **Fallback rate:** How often the interaction ends in a human handoff, a canned response, or an abandoned task

The types of metrics listed previously help organizations measure how the customer experienced the interaction with an AI agent. They help inform the patterns that emerge from customer interactions. They help track important shifts in the agent-customer interaction from intent to sentiment and the frustration that occurs when customers must repeat context the AI agent should have already known.

Experience signals are critical to improving AI agent behavior, including agent decisions that stall customer interactions and create friction. They provide insights that help align customer behavior with AI agent behavior. And they help apparel, home goods, and travel and hospitality organizations protect margins by reducing redundant, circular customer interactions that frustrate customers and waste expensive tokens.

II. Definitions

- **AI agent** is a software entity that perceives its environment, reasons about goals, and takes autonomous action to achieve outcomes — without step-by-step human instruction for each task. The defining trait is not "using AI" but autonomous goal pursuit where the

agent decides the sequence of actions needed to reach an objective rather than executing a prewritten script in resolving a customer inquiry or service request.

- **Agent experience** is the customer's perceived outcome and efficiency of an agent interaction: whether they got what they came for, how efficiently, and whether they came away satisfied. It is derived from behavioral signals rather than self-reported feedback.
- **Agentic AI** is software that uses machine learning and deep learning techniques to exhibit agency — setting goals, making decisions, and taking actions through the perception-reasoning-action loop. Agentic AI software "exhibits agency" by executing that three-stage loop — perceive the environment, reason about goals and options, and act to pursue an outcome.
- **AI agent evaluations** include a set of tools and methods used to test the performance, accuracy, fairness, and other qualities of agentic AI systems, typically in advance of shipping an agent, workflow, or agentic capability to production. Since AI agents can act in real time, the trend is toward "continuous evaluations" in production.

III. Benefits

Comprehensive measurement of AI agent performance will enable organizations to drive the metrics that really matter to the business, such as customer experience, loyalty, retention, and customer lifetime value.

Simple evaluations provide very little insight into the overall interaction and mostly provide an inside-out view of the transaction, even when evaluations run continuously. The voice of the customer is lost in the drive toward technical optimization of performance metrics. Evaluation and observability vendors are adding user-sentiment and frustration features.

Measuring both technical performance and customer experience provides enterprises with the ability to understand the trade-offs between cost of service and customer experience, which is critical for any enterprise customer service operation. It allows the CX group to route inbound inquiries to the rightsized models, provide the appropriate level of reasoning, and design escalation paths for more complex inquiries where deeper reasoning may be required.

Experience inefficiency and token waste are the same problem viewed two ways. Every unnecessary clarification or retry is both a worse experience and a higher bill. Measuring experience by segment shows why the agent failed, such as a missing tool parameter, stale context, or an unclear instruction. Most of these issues can be fixed without changing the model.

IV. Trends

While much of the attention has been on the automation of CX, the true strategic value is measured by whether the AI system is customer centric: whether the use case improves outcomes for customers and consumers.

The barrier to AI agent autonomy is the CX "blind spot": the missing data and observations that give enterprises confidence that the customer experience was positive. "Thumbs-up," answer relevancy, accuracy, and resolution, as currently measured, don't necessarily translate into customer satisfaction, loyalty, and retention. That is partly why agent autonomy in customer-facing AI agents has stalled.

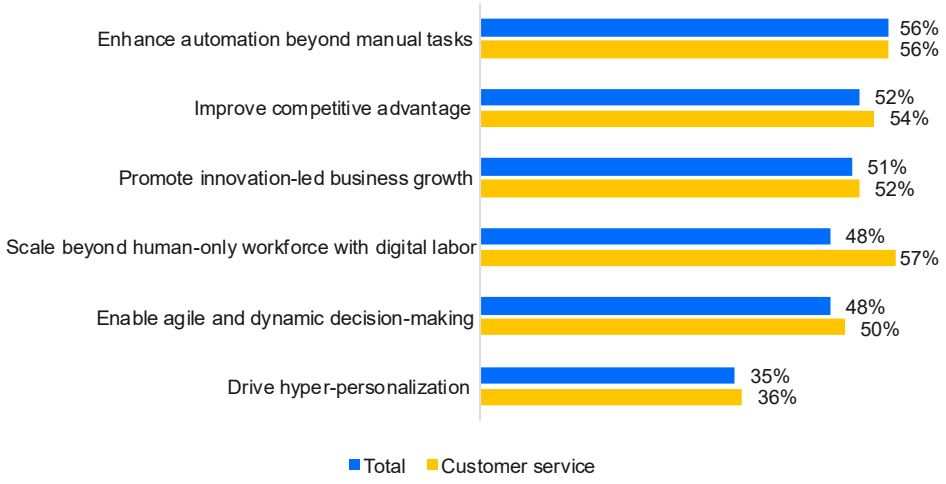
A secure and compliant interaction that delivers a poor customer experience is a failure. And it has important cost implications. Spikes in token spend can look like the AI agent is working harder to solve the customer's problem, but it could be that the agent is failing. Current evaluation platforms that don't tie token spend to CX metrics can hide completely wasted token costs.

This is especially important since the use cases where agentic adoption is happening fastest include customer service, customer support, sales, and commerce. Failing to measure and understand the customer experience with commerce and shopping agents, for example, can mean lost sales and higher token costs. Knowing that a cart was abandoned is important telemetry. Knowing "why" is even more crucial to finding and fixing issues. Enterprises cannot diagnose the issue and get to the root cause unless they move beyond efficiency metrics and prioritize customer experience metrics measured and compared across segments of the entire population. Enterprises need to prioritize CX (see Figure 1).

Figure 1

Customer service leads in digital labor ambitions; customer experience is absent from the priorities

Q. Which strategic priorities are most critical for current agentic AI initiatives being done or considered within the customer service function?



57% of customer service leaders prioritize scaling beyond a human-only workforce, the highest of any function. None of the listed priorities measures whether customers were well served.

n = 1,967 for total, n = 148 for customer service

Source: IDC's *Early Adopter Agentic AI Functional Use Case Study*, March 2026

Enterprises are not to blame for the misplaced focus on technical measures. Given that most agentic implementations have been driven by IT up to this point, evaluation platforms have focused on the metrics most important to IT versus the line of business. IT focuses mostly on execution, while marketing and sales professionals focus on customer experience.

To attain more business-oriented objectives, such as increasing conversion rates online, enterprises are seeking more complete evaluations to provide more autonomy to agents in customer interactions. The lack of complete understanding of impact on customer experience inhibits full autonomy. Customer service organizations lag behind other functions (61% versus 68%) in agent autonomy (see Figure 2).

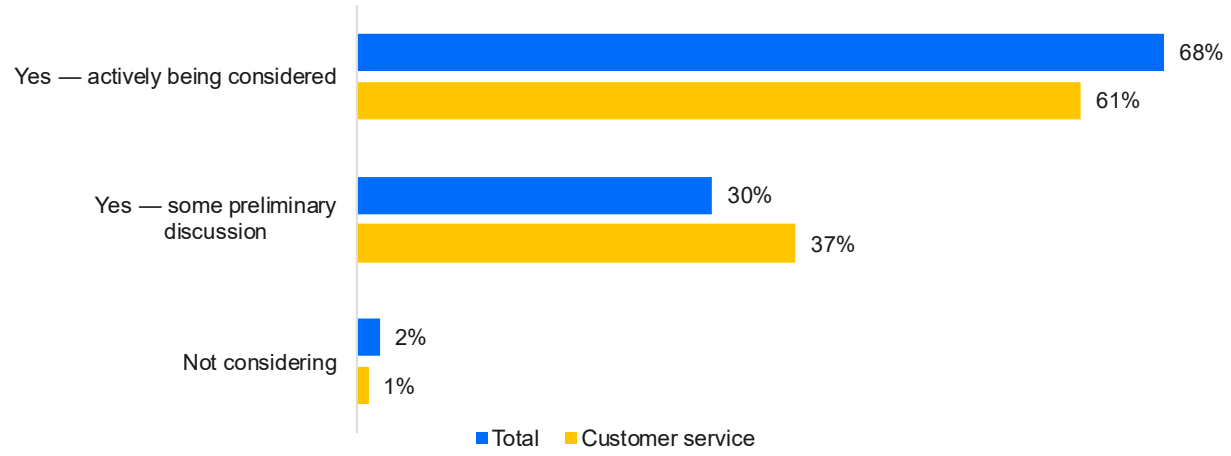
As we continue to scale agents across the enterprise and increasingly deploy teams of agents, the ability to measure agent behavior with traditional evaluation platforms and observability metrics has fallen behind. While new approaches like LLM as judge and other advanced methods help scale technical evaluations, they do not provide insight into the customer experience. LLM judges read only the transcript. An AI agent, for example, can produce a fluent, polite final message after having called the wrong API, silently failed a tool call, or looked up the wrong account. Again, what looks like a successful interaction in the transcript could be masking failure from the customer's perspective. Customer experience requires a more robust evaluation. LLM judges can answer, "Was this response correct?" but the transcript is silent on the more important question: "How did this work for the customer on the other end?"

As enterprises experienced significant cost overruns in their agentic AI budgets, token economics and AI cost management began to dominate the conversation. Enterprises are looking for solutions that help them align token costs with customer requirements to avoid overspending. In fact, two-thirds (67%) of enterprises ran over their agent spend budget in the past 12 months (source: IDC's *Future Enterprise Resiliency and Spending Survey [FERS Survey]* Wave 4, July 2026). Without granular data that enables enterprises to match spending to customer experience outcomes, enterprises are flying blind about the impact of that spend. The fastest way to cut agent spend is to eliminate the turns customers never needed.

Figure 2

Customer service lags in AI agent autonomy

Q. Are there discussions within your organization to allow AI agents to have more autonomy within customer service?



n = 1,967 for total, n = 148 for customer service

Source: IDC's *Early Adopter Agentic AI Functional Use Case Study*, March 2026

V. Technology or Vendor profile

Conviva is a real-time quality-of-experience (QoE) analytics platform. Founded in 2006, the company has focused on QoE analytics for streaming. Its approach is intended to extend streaming providers beyond quality-of-service metrics, such as whether a stream started, to measures of whether viewers actually had a good experience. According to Conviva, its platform has supported 10 Olympic Games, 20 Super Bowls, and 5 FIFA World Cups. During the opening week of the 2026 FIFA World Cup, it supported more than 20 streaming providers at 70 million peak concurrent plays. Customers include Fox, Paramount+, DAZN, MLB, and Sling.

Conviva is applying the same discipline to AI agents. Conviva is designed to measure the impact on the customer, including behavior on the surrounding website or app that never appears in the transcript. The platform is built to analyze sessions and conversations to identify the behavior patterns associated with customer experience, token costs, and revenue.

Conviva organizes its approach around four principles, which it calls EPIC:

- **Experience: Stateful QoE metrics, inside and outside the conversation.** In video, Conviva converted raw playback events into metrics tied to viewer experience, such as startup time, rebuffering, and picture quality. It is applying the same approach to agents, with experience metrics designed to be computed from behavior both inside the conversation and on the surrounding website or app. Inside the conversation, they are intended to include clarification rate. Outside it, they are intended to include manual

searches mid-conversation, purchases made outside the chat, downstream rework, and return visits about the same issue. Because these signals unfold over time, Conviva uses a Time-State analytics engine designed to track the state and sequence of each session rather than score isolated turns.

- **Population: Full-census client-side telemetry.** Conviva is designed to collect telemetry from all sessions and devices rather than sampling. Instrumentation is intended to live in the application, browser, device, or agent, capturing customer behavior rather than inferring it from server logs or transcripts. Failures that affect a small segment often surface only across the full population.
- **In production: Real-time automated insights.** The platform is designed to analyze live traffic as it happens rather than curated test sets. It is intended to baseline KPIs automatically, detect anomalies, run root cause analysis, and rank insights by business impact, so teams can learn about problems from the data rather than from their customers.
- **Contrastive: High-dimensional pattern detection.** AI-driven analysis is designed to compare overperforming and underperforming segments. This approach is intended to isolate the segment where experience breaks down and help explain why.

Challenges

Because this level of analysis is new and offers a different approach to AI agent performance analysis, Conviva faces some challenges:

- **Category education.** Agent experience is an emerging discipline. Conviva must help buyers distinguish behavioral experience measurement from both traditional CX surveys and evaluation scores. Conviva must help enterprises widen the aperture on how customer experience metrics can be included in the evaluation of AI agent performance. Robust multidimensional analysis of customer experience and agent behavior is new to enterprises. These skills have largely been siloed, with IT owning technical performance and marketing owning customer experience. Conviva will need to help bridge the IT-versus–line of business (LOB) divide.
- **Split ownership.** Agent programs are often divided between IT and AI engineering on one side and business owners on the other. Conviva must engage technology leaders and line-of-business leaders at once.
- **Instrumentation.** Capturing behavior outside the transcript requires client-side data collection beyond trace ingestion, which adds deployment steps.
- **Competitive movement.** Evaluation and observability vendors are adding user-sentiment and frustration features. Conviva must demonstrate the depth and scale of its behavioral analysis.
- **Brand perception.** Conviva is well known in media and streaming. It must build equivalent credibility with enterprise AI buyers in other industries. Many enterprises may not be aware that the current shortcomings of evaluations have already been addressed in other industries and CX domains.

VI. Conclusion

IDC believes the only path to unlocking the full value of autonomous AI agents in customer-facing interactions is to ensure that organizations have insights into customer experience. For most organizations today, it remains a "blind spot" in the evaluation of AI agents and agentic AI systems. The most direct path to optimizing agent performance is to see both sides of the agent interaction: how it performed from an inside-out perspective and how the customer experienced the interaction. Further:

- **Measure experience, not just responses.** The response shows whether a single output in the interaction was "correct." It does not reveal whether the customer's actual problem got solved. An AI agent response can be totally accurate and still be a bad customer experience.
- **Measure the full population in production, not samples.** Many current evaluation methods rely on sampling. This approach can systematically miss CX-specific failures by reporting that the "average" interaction went fine, leading to unseen erosion of customer loyalty and lost revenue.
- **Compare segments to find where the agent fails.** A single number, such as "resolution rate," can mislead. It misses significant and costly patterns that averages wash out. AI agent failures are rarely uniform. They often concentrate in specific customer segments. Enterprises can only know "where and why" CX failures occurred by comparison across the entire population.
- **Tie experience metrics to token spend**. Token spend only measures the cost of running the agent. A spike in token cost could mean it's doing more valuable work to solve the customer's problem, or that it is failing — vastly different outcomes. High token spend with poor customer experience is pure waste. The only way to know is by tying token spend to CX metrics. Doing so helps expose inefficient patterns, such as AI agents spending the first couple of minutes with customers gathering context, which is a token-spend and CX problem simultaneously. Tokens and customer patience are both spent on the same wasted motion.

Without a complete picture, when it comes to customer experience, the current state of agent performance evaluations is at best inferred, and at worst a guess.

About the IDC analyst



Tim Law
Research Director, AI and Automation

Timothy Law is a research director for AI and Automation, responsible for the generative AI life-cycle tools and technologies research practice. Mr. Law's core research coverage includes the evolution of generative AI infrastructure and platforms, foundation models, developer tools, observability solutions, agentic systems, and generative AI services. This research analyzes the trends and developments in the AI software markets, including the costs, benefits, and impact of generative AI technologies.

More about Tim Law

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.

One Beacon Street, Suite 33100, Boston, MA 02108, USA

T +1 508 872 8200

blogs.idc.com | LinkedIn @IDC | www.idc.com

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community make fact-based technology decisions and achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. CCPA