CONVIVA

Outcome-Based Agent Evaluation and Optimization

Lessons from Pioneering Video Streaming Delivery and Optimization Position Conviva to Power the Agentic Economy

Businesses continually search for ways to reduce the friction consumers face when using their services — from the early age of telephone-based interactions to early web applications, to modern cloud-hosted e-commerce and mobile experiences.[1] The emergence of AI-based agents is the next natural step — acting as intermediaries for consumers and interacting with other agents. We are now at this exciting inflection point where agents can reshape customer interactions and potentially disrupt the entire digital economy — dramatically improving efficiency and user experience.

The Billion Dollar Catch: Performance of Agents in Production Is Unpredictable!

While there is immense potential to improve business outcomes, using agents could be a massive waste of investment, damage brand reputation, or erode consumer trust.

Let's explore two incidents to illustrate how AI agents can malfunction and why their real-world performance may fall far short of expectations.

Imagine you are the CTO of a popular e-commerce storefront. You are excited to roll out a new conversational shopping assistant to simplify the customer experience. Your AI/ML and engineering teams have worked hard to build a proof-of-concept and have tested it in the lab. You believe it will help customers:

- Discover relevant products more easily
- Compare different options side-by-side
- Access comprehensive product details and specifications
- Get up-to-date pricing and shipping information

Yet, even with the best preparation, unforeseen problems can arise resulting in low ROI of your agent, bad business outcomes, and a possible erosion of customer trust due to failed interactions.

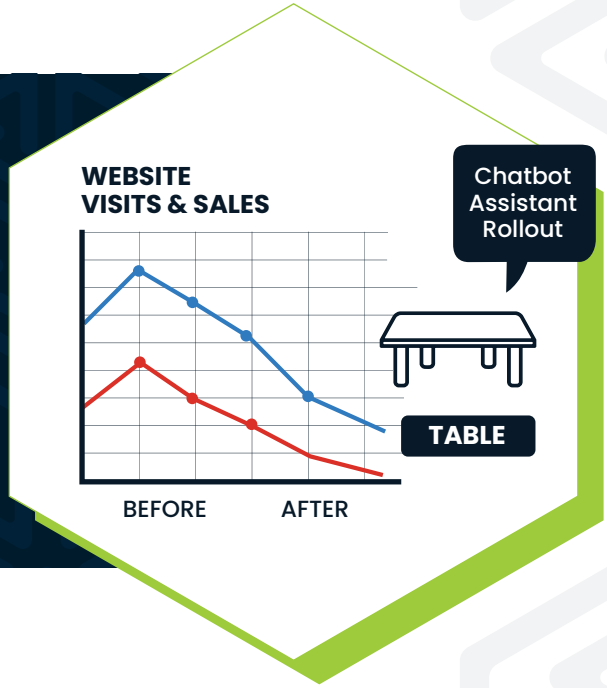
1. *https://arxiv.org/abs/2505.15799 The Agentic Economy*

Scenario 1

Incomplete product recommendations → Revenue loss

After rolling out the agent, analytics teams discovered that click-through rates and sales for high-value products dropped in certain regions. The AI models and prompts may have had blind spots that prevented the agent from capturing and recommending higher-value items to users.

What's needed: Teams need to analyze end-user sessions by examining both agent interactions (such as user queries) and consumer digital behaviors (including link clicks, page navigation, and rage clicks). This analysis should automatically identify patterns connecting user experiences with session attributes and business outcomes. For example, certain query types may be poorly handled, leading to reduced conversion rates.



Scenario 2

User conversion failure due to misinterpretations and data errors

Just as incomplete product recommendations can erode revenue, another common failure mode is when agents misinterpret customer intent or provide inaccurate data, directly undermining conversions.

Consider a price-conscious customer searching for a specific product. They want to compare similar items in their price range, read relevant reviews, and make an informed purchase decision. Unfortunately, the agent fails to provide accurate product comparisons, suggesting products outside the customer's budget or missing relevant products with recent price drops. After trying several queries and manual searches, the customer abandons the site without making a purchase. Several factors could cause this failure: the agent misinterpreting customer intent, failing to access updated pricing databases, or incorrectly processing backend responses.

What's needed: Teams must analyze customer sessions by examining agent interactions (such as price-comparison queries or repeated questions) alongside digital behaviors (including whether customers click on agent-provided links). This analysis should automatically identify patterns linking customer experiences with session data, high-dimensional contextual attributes, and business metrics. For example, repeated queries with no clicks may indicate poor agent responses and lead to reduced click-through rates.

The unreliability of agents and AI-assisted solutions in real-world settings is a well-documented problem, acknowledged by academics and practitioners.[2] [3] Early reports suggest we are in the early stages of this agentic revolution even as most pilots are failing.[4] These failures can stem from multiple sources: unclear customer intent, AI model errors and hallucinations, planning problems in agent systems, or unclear input-output behaviors in workflow steps. Third-party services outside the company's control, including model providers, agent protocols and server caching, compound these issues.

Why Standard Monitoring and Observability Tools Fall Short for AI Agents

Traditional and agent-based applications may seem similar at first. Both process user requests and produce outputs. This similarity might suggest that conventional backend observability tools (such as Datadog) or product analytics platforms (such as Amplitude or Quantum Metric) would suffice. Alternatively, you might consider supplementing these tools with AI benchmarks or large language model-based testing from emerging vendors such as LangChain.

These tools may be a good fit for organizations focused on ML observability and explainability rather than customer outcome optimization. They are best suited for ML and data science teams who need drift detection, bias monitoring, and analysis for production models. They are not optimal for enterprises that must tie end-user flows and outcomes directly to revenue or customer journeys.

Conviva is designed for this gap — delivering outcome-based performance analytics for agents that directly optimize customer experience, AI agent performance, and business results.

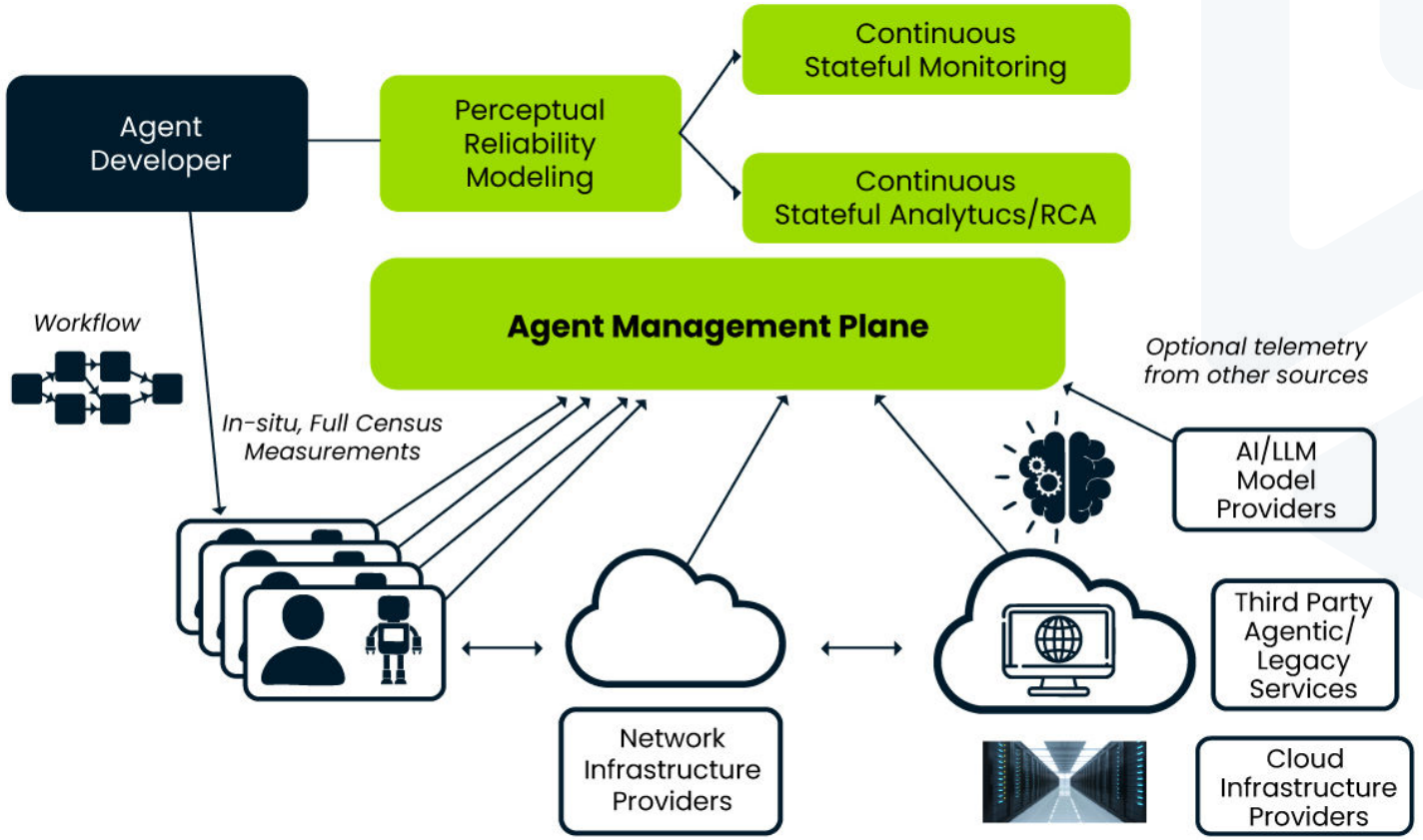
To understand the difference, consider how agentic systems diverge from traditional applications.

1. Traditional applications handle deterministic requests, such as database queries or webpage requests. Agent workflows process inherently ambiguous intents that can lead to complex, unpredictable interactions between users, agents and applications.
2. Classical applications execute deterministic data and computing actions. Agentic workflows could have non-deterministic behaviors from agents or model providers.
3. The notion of "success" of the workflow in traditional apps is restricted to simple performance or error metrics (e.g., request success, latency, HTTP 404).

In agentic workflows, we need to explicitly consider reliability as measured by patterns and milestones of the customer journey. When traditional services do interact with third-party services, these interactions tend to be few and scoped. Agentic systems may involve more third-party services (e.g., LLM providers, other agents) with non-deterministic behaviors, that can induce cascading issues.

Conventional approaches have fundamental limitations for agent deployments. Backend observability tools that focus only on server metrics, logs, and traces miss critical interactions. Offline benchmarks and testing methods, including large language model evaluations, fail to capture real-world complexity. Traditional monitoring solutions that sample only endpoint interactions cannot observe the full range of user-agent-application interactions or connect them to business outcomes.

2. https://arxiv.org/abs/2503.13657 Why Do Multi-Agent LLM Systems Fail
3. https://nextbigteng.substack.com/p/the-state-of-ai-2025 The State of AI 2025
4. MIT Says 95% Of Enterprise AI Fails https://www.forbes.com/sites/jaimecatmull/2025/08/22/mit-says-95-of-enterprise-ai-failsheres-what-the-5-are-doing-right/



Successful agent deployments require more than backend AI models, prompt engineering and offline testing. Teams need a comprehensive management system to continuously improve agent workflows. This system must be powered by *full census* measurements *of perceptual performance* collected *in-situ* by *the user-facing endpoints.*

By *perceptual performance,* we mean the need to track the semantic outcomes of workflows rather than low-level metrics such as time to first token or latency. Such outcome-centric perceptual performance measures can be direct, such as binary task completion, or indirect, such as inferring user frustration.

By *in situ*, we mean that: (1) the agent endpoint coordinating the various subtasks is in the best position to provide the contextual information to compute the perceptual reliability measures and (2) these measures are made in production, or "in the field," rather than "closed world" offline tests.

	Traditional Observability	LLM Observability	Conviva
Primary Use Cases	• **DevOps** • **Agent Uptime**	• **LLM QA** • **Prompt evaluation** • **Offline tuning**	• **Agentic system reliability** • **Customer trust** • **Business growth**
Scale	Built for Infrastructure scale (thousands of server instances)	Single or small sets of model calls (evaluation-scale)	Designed for autonomous agent sessions at any scale (up to billions)
Depth: Scope of Monitoring	Component-level (servers, APIs, infrastructure)	Isolated LLM prompts & responses	End-to-End: user → agent → system → business outcome
Semantics	Low semantic depth: no sentiment or conversation context	Captures sentiment and local conversation context (per prompt)	Captures full sentiment and context across non-deterministic conversations and sessions
Insights Provider	App & infrastructure health	Model-centric metrics (e.g., latency, hallucination)	Outcome-based agentic system performance
Dimensionality	Static, system-level health metrics and anomalies	Prompt analysis, hallucination trends (static model dimensions)	High-dimensional behavior + sentiment analysis across entire session

Finally, to capture the long tail of possible interaction - and outcome-related failures, we argue for full census measurements. In other words, randomly sampling a few users (as in traditional RUM) or a few interactions is insufficient for modeling perceptual reliability or diagnosing when and why agents fail. Instead, we need to capture events relevant to perceptual reliability outcomes for every agent and across all its actions and subtasks, rather than relying on a small subset.

Why Conviva Is Uniquely Positioned to Help Businesses Succeed in the Agentic Economy

In many ways, we see natural parallels between the emergence of the agentic economy and our pioneering work in the history of the largest (until now) "killer app" on the Internet — Over-the-Top, or OTT, Internet Video. History repeats itself!

Requirements	Conviva Innovations in Internet OTT Video	Success for Internet of Agents
Perceptual Performance Monitoring	Conviva pioneers the case for quality of experience	We need perceptual performance measures to evaluate agents, seen in the context of the digital application and user interactions, not just latency or text judging
In-Situ Monitoring	Conviva pioneers video and application "sensors" for auto collecting relevant telemetry	We need lightweight but robust endpoint sensors to capture interactions among the user, the agent, and the digital application footprint
Full Census Monitoring	Conviva´s breakthrough technology, the Time-State platform enables stateful context-aware analytics and AI capabilities for Internet-scale analysis	We need scalable high dimensional analytics to automatically surface opportunities for improving Agents' performance "in the field"

Conviva pioneered the case for moving "measurements up the stack" from low-level quality of service (QoS) measures to user-perceived quality of experience-centric (QoE) measures as the key drivers of business success. Conviva brings decades of experience in helping Internet-scale OTT providers define, measure, and optimize the relevant perceptual quality of experience to improve their business outcomes.

The history of OTT Internet video taught us that the endpoints are in the best position to obtain the necessary telemetry to model their performance and reliability requirements. Conviva's deep experience in developing and deploying lightweight endpoint sensors at massive scale — integrated in billions of diverse endpoints — enables automatic collection of the relevant telemetry needed to power business outcomes. In the agentic era, this is critical to capturing the diverse interactions between the users, the agents, and digital applications.

Successful streaming applications adopted application-layer management to work around the unpredictability of the infrastructure, by continuously identifying and analyzing opportunities for improving business outcomes at massive scale and dimensionality. The Conviva Operational Data Platform with patented Time-State Technology processes trillions of events from billions of endpoints worldwide, automatically identifying insights and improvement opportunities for both real-time and long-term optimization.

These advances are not just about scale or efficiency — they are about creating a foundation of reliability and trust. For the agentic economy to thrive, systems must consistently transform raw data into outcomes that people can depend on, ensuring that every interaction feels both credible and helpful to the customer.

The future of the agentic economy depends on whether businesses can deliver **trustworthy, outcome-based agent experiences**. Success isn't about how fast an API responded, or if an agent produced a grammatically correct sentence. **Consumer impact is the only metric that matters**. You must measure the quality of the conversation and the outcome from the user's perspective, not backend reliability. This is quality of experience (QoE) for agents.

Success is whether the **consumer felt their goal was achieved** — because in the end, customer perception is reality.

That's why four truths must guide this new market:

1. Perceptual end-user outcomes in production are the only objective way to measure agent success.
2. Non-determinism of agents and their interactions with other third-party entities cannot be captured offline or by sampling.
3. Monitoring outcomes and understanding why failures occur requires more than reviewing a chat log or backend servers alone. A holistic view across the agent, the application, and the backend is needed to measure both the perceived outcome and the reasons for failure.
4. Connecting the dots requires stateful, contextual analysis within both the session or conversational flow and the system or AI state.

Together, these truths demand a new kind of analytics — one that is **stateful, contextual, and Internet-scale.** Sampling, offline tests, or siloed observability tools won't suffice. What's required is **continuous, in-the-field visibility** into every agent interaction, at full census, across all dimensions of the digital experience. Conviva is the only platform with the compute power to deliver deep dimensional insights at scale.