AI Next

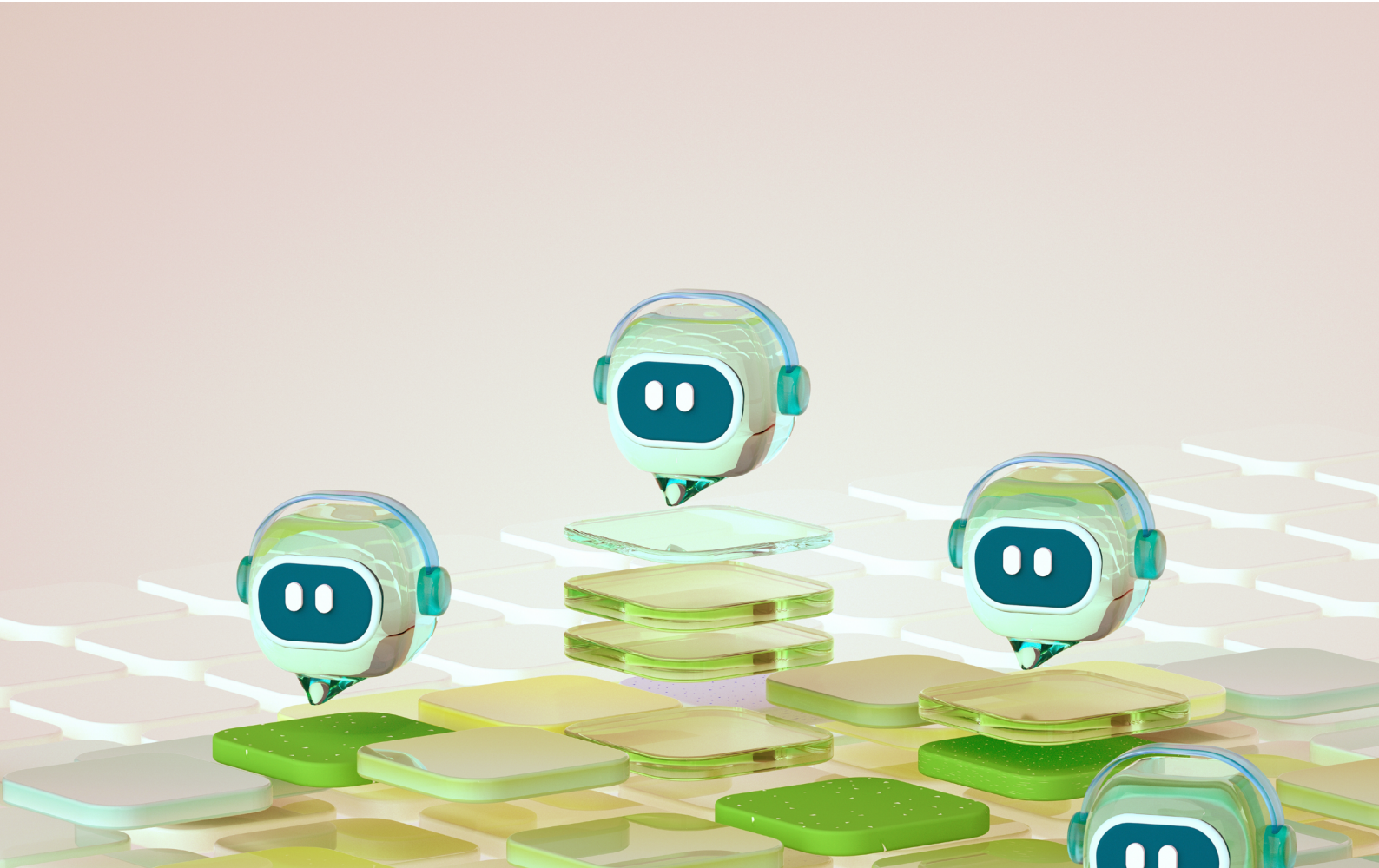
edgeverve
An Infosys company

# Enterprise AI Agent Evaluation:

## The Platform-Based Approach to Success

## Table of Contents





# 1. Abstract

As AI systems evolve into more complex agents capable of sophisticated reasoning and decision-making, the evaluation challenge becomes both more critical and significantly more complex. Traditional, static metrics are no longer sufficient in dynamic, real-world environments; instead, agentic systems require richer, multilayered evaluation approaches.

**This paper highlights that:**

Evaluation is no longer a peripheral hygiene practice but the foundational operating system for agent development.

Agentic AI must be assessed holistically across inputs, intermediate actions, and final outcomes.

Achieving this level of rigor at scale requires a platform-based approach.

We present a comprehensive evaluation framework, outline methodologies tailored for agentic AI, and demonstrate how a unified platform simplifies the orchestration of multilayer evaluations, making rigorous, scalable, and continuous assessment practical and accessible.

# 2. Why Evaluations Matter—and Why They're So Hard with Agentic AI

Enterprises are rapidly adopting AI agents to automate multistep workflows and accelerate ROI. While many have begun experimenting with agentic systems, a growing number are now progressing into full-scale deployment. As these systems mature, leaders expect clear evidence of reliability, safety, and measurable business impact. This makes rigorous evaluation frameworks essential to reduce risk, ensure consistent performance in real-world environments, and justify continued investment.

However, evaluating agentic AI is fundamentally more challenging than traditional ML or deterministic systems. Several characteristics of agents introduce new layers of variability, ambiguity, and hidden failure modes:

**Non-deterministic paths:**
Decision-making modules may select multiple valid reasoning paths or tool call sequences to answer the same request. This variability breaks the assumptions of conventional evaluation frameworks, which are built around fixed inputs and predictable execution flows.

**Increased failure points across multi step workflows:**
Agent workflows span multiple steps, tool invocations, API calls, and parallel branches. Failures can occur at any stage, and meaningful evaluation requires visibility into the entire session, not just the final turn.

**The "hidden failure" problem:**
Agents may produce a correct final answer through an incorrect, unsafe, inefficient, or suboptimal process. These failures remain hidden if evaluation focuses solely on outputs rather than the reasoning path that produced them.

Because of these challenges, evaluation must extend far beyond checking whether the final output is correct. It requires a multidimensional assessment, including:

- Did the agent choose the appropriate tools and actions at the right time?
- Was the agent's reasoning coherent, safe, and aligned with policy constraints?
- Did it manage uncertainty, fallbacks, and error states responsibly?
- How robust is its behavior across seeds, user variations, and scenario changes?

In short, agent evaluation must capture not only **what** the agent did, but **how** and **why** it did it.

# 3. Where Agentic AI Evaluations Fit in Your Process

Evaluation is not a single gate at the end of development—it is a continuous discipline that spans the entire agent lifecycle. The diagram below illustrates how evaluation activities map across the **Plan → Build → Test → Deploy → Monitor** pipeline, with pre-deployment evals, production evals, and governance operating as an integrated layer throughout.

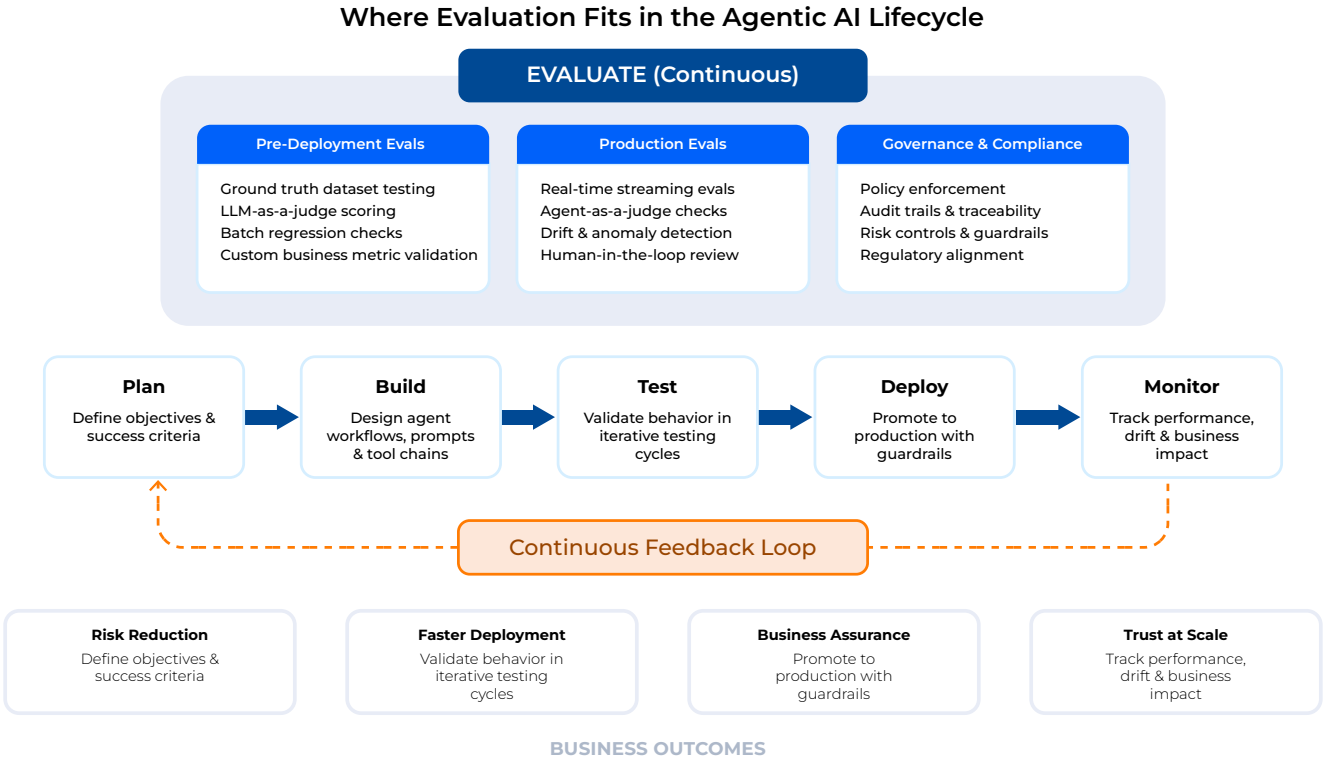



Fig. 1: End-to-end view of how evaluation integrates into the Agentic AI development process

# 4. Devising the Evaluation Strategy

Given the inherent complexities of evaluating agentic systems, the most effective place to begin is with absolute clarity—specifically, a well-defined understanding of what success looks like for the agent. Rigorous evaluation is only possible when that definition is explicit, unambiguous, and directly linked to measurable outcomes.

An effective evaluation strategy therefore starts with a precise articulation of the agent's purpose. The core question is simple yet foundational: What does success mean for this particular agent? These success statements must be concrete enough to translate into testable metrics and operational expectations.

Once success is clearly defined, it can be evaluated holistically across four pillars, each capturing a different dimension of agent performance:


### 1 Comprehensive Agent Quality

This pillar looks at the agent's overall performance in real interaction scenarios—how effectively it completes tasks, how coherent and grounded its responses are, and how well the experience aligns with user expectations. It mirrors production behavior and functions like an integration-level assessment of the entire interaction flow.

### 2 Process Flow & Reasoning Trace

Beyond the final answer, this pillar examines the quality of the agent's internal reasoning. It assesses whether the agent takes logical, efficient steps, chooses tools appropriately, and maintains a sound decision-making path throughout multistep workflows. This is essentially a deep dive into the chain-of-thought execution pattern and operational correctness.

### 3 Trust, Safety, and Robustness

This dimension focuses on how well the agent handles unexpected, adversarial, or edge-case situations. It evaluates the agent's resilience to manipulation, its ability to fail safely, and its consistency in avoiding unsafe or biased behavior. The goal is to ensure dependable performance even in nonideal conditions.

### 4 Operational Performance and Scalability

This pillar evaluates whether the agent can operate effectively under real production constraints—handling scale, maintaining low latency, controlling costs, and remaining reliable over long-running workloads. It ensures not just intelligence but operational readiness.

# 5. The Multilayered Approach to Agentic Evaluation

Building on the pillars and success criteria defined earlier, the next step is selecting the right evaluation methodologies—because even the strongest strategy fails without reliable data and rigorous measurement techniques. And at the heart of every evaluation method lies its foundation: the quality of the data it runs on.

High-quality ground truth datasets are essential. They pair inputs with expected outcomes and—particularly for agentic AI—often capture expected intermediate behaviors as well.

To establish this foundation, some of the key components include:

- Creating comprehensive and representative test cases
- Managing and versioning datasets to ensure traceability and repeatability

## Evaluation Methodologies for Agentic AI

### Predefined Scorers

Platforms often include pre configured scoring functions optimized for common evaluation requirements. These provide a fast and consistent baseline, enabling teams to begin evaluations quickly and then extend or tailor them with custom logic as needed.

### LLM-as-a-Judge

LLM-as-a-Judge as an evaluation approach where a large language model (LLM) is used to score or review different components of an AI system. The methodology involves prompting a capable LLM to assess the quality of a wide range of outputs, including those produced by other models or by human annotators. This approach becomes especially valuable when statistical comparisons against ground truth are insufficient or not feasible—such as when no ground truth exists, or when working with unstructured outputs that lack consistent, reliable evaluation metrics.

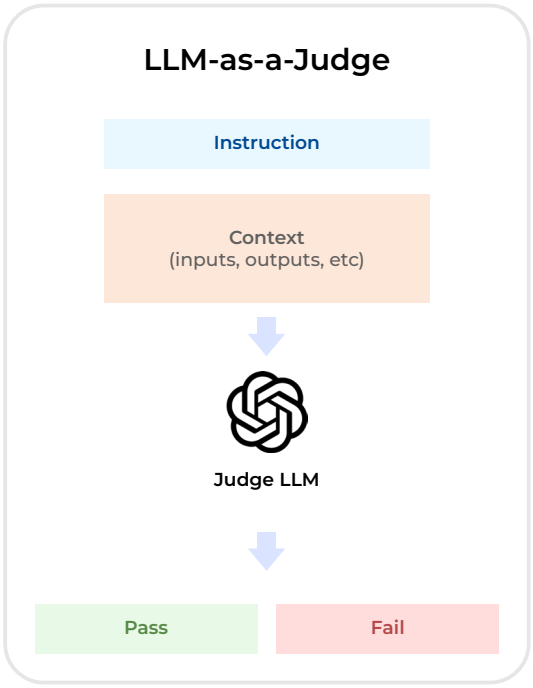



Fig. 2: LLM-as-a-Judge workflow (Source: ML Ops documentation)

## Agent-as-a-Judge

Agent-as-a-Judge represents a paradigm shift in LLM evaluation. Instead of simply assessing inputs and outputs, these judges act as autonomous agents equipped with tools to investigate the application's execution in depth.

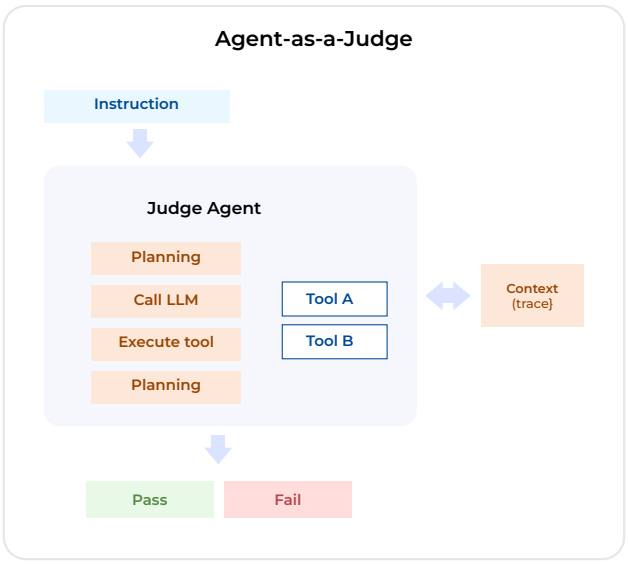



Fig. 3: Agent-as-a-Judge workflow (Source: ML Ops documentation)

## Custom Metrics

Custom scorers offer maximum flexibility, allowing organizations to define evaluation metrics tailored to their unique business context. These metrics can be based on simple heuristics, domains- specific rules, structured scoring rubrics, or advanced programmatic logic. They are essential when evaluating outcomes tied to enterprise-specific workflows, compliance requirements, or business KPIs.

## Human Evaluation Method

Human evaluation remains a critical part of any robust evaluation strategy. Human judges bring domain expertise and contextual understanding that automated methods often miss. They help establish ground truth, validate whether automated scoring aligns with real-world expectations, and uncover nuanced failure modes that reflect actual user experience and business impact.

## Comparison of Different Methodologies

| Method | What they evaluate | Speed | Cost | Use cases |
|---|---|---|---|---|
| **Predefined Scorers** | Accuracy, similarity, safety, toxicity, and standard quality metrics | Fast (deterministic scoring) | Lowest | OCR & Document Extraction, Toxicity / Safety Filtering, Search & Retrieval Quality, Automated Reporting |
| **LLM-as-a-Judge** | flexible, scalable alternative to human evaluation, evaluating response quality, factual accuracy, safety, bias, and any other requirements specified by prompt | Moderate (depends on LLM size, prompt depth, and reasoning length) | Moderate to High — driven by LLM inference cost, prompt complexity, and number of evaluation runs | Classification Tasks, Customer Support QA, Document Summarization, RAG Pipeline Evaluation |
| **Agent-as-a-Judge** | Complete tool-execution traces and trajectory, intermediate reasoning steps | Slower | High (more context and tool usage) | Loan or Credit Decisioning, Claims Processing Workflows, Policy Analysis, Code Reviews |
| **Custom Metric** | Business-specific rules, domain constraints, KPI alignment, policy compliance | Medium (depends on rule complexity) | Variable (depends on logic + compute) | Underwriting & Eligibility, Regulated Industry Workflows |
| **Human Evaluation** | Subjective quality, nuance, safety, tone, high stakes judgment | Slowest | Highest (human time + effort) | Compliance-Critical Decisions, Healthcare & Medical Support |

## 6. From Development to Production: Continuous Evaluation

With evaluation methodologies established, the final step is ensuring they operate as a continuous part of the agent lifecycle rather than a one-time checkpoint. By integrating the evaluation framework directly into the engineering workflow, organizations can operationalize evaluation as an automated, ongoing process—one that continuously monitors, scores, and validates agent behavior. This closed feedback loop ensures that the agent becomes progressively more accurate, resilient, and reliable with every iteration and update.

## 7. The Platform Imperative

The evaluation methodologies, data foundations, and lifecycle practices outlined so far are necessary but not sufficient on their own. In isolation, they remain manual processes—difficult to scale, hard to reproduce, and impossible to govern consistently across an organization deploying multiple agents.

This is the **platform imperative:** the recognition that rigorous, continuous evaluation at enterprise scale requires purpose-built infrastructure—not a collection of point tools stitched together. The difference is structural. A platform approach means evaluation is not something that happens outside the agent development process; it is woven into how agents are built, tested, deployed, and monitored.

Consider the alternative. Without a unified platform, enterprises typically face a fragmented landscape: one tool for tracing, another for scoring, a separate system for governance, and manual processes to bridge them. Each tool may be excellent in its domain, but the gaps between them—the missing audit trails, the disconnected feedback loops, the inability to enforce consistent evaluation criteria across development and production—create exactly the blind spots that agentic evaluation aims to eliminate.

A platform-based approach resolves this by providing a single operating layer that encompasses the full evaluation lifecycle: curating and versioning ground-truth datasets, orchestrating multi-methodology evaluations (from automated scorers to human review), enforcing governance and compliance policies, capturing end-to-end traces for observability, and closing the feedback loop from production insights back into development. When these capabilities are unified, evaluation becomes a continuous, automated function rather than a periodic, manual effort.

The question for enterprise leaders is therefore not whether to evaluate—that is a given—but whether their evaluation infrastructure can keep pace with the complexity and scale of their agentic deployments. The next section examines how **EdgeVerve AI Next** operationalizes this platform vision.

# 8. Evaluating at Scale with EdgeVerve AI Next

## The Industry Challenge: Fragmented Tooling, Limited Visibility

The market for AI evaluation tooling has grown rapidly, with many vendors offering specialized capabilities in observability, scoring, and guardrails. These tools have advanced the state of the art considerably—introducing concepts like evaluation foundation models, distilled judge models for low-latency production monitoring, and automated eval-to-guardrail lifecycles. However, most of these solutions serve a specific slice of the evaluation problem.

The result: enterprises are left stitching together multiple point solutions, creating integration overhead and leaving critical gaps between development-time testing and production-time monitoring.

For enterprises operating at scale—where agents handle sensitive financial workflows, regulated processes, or customer-facing decisions—this fragmentation is untenable. What is needed is a platform that unifies agent orchestration, evaluation, observability, and governance in a single operating layer, so that evaluation is not an afterthought bolted on at the end but a continuous, embedded function of how agents are built, deployed, and monitored.

## How EdgeVerve AI Next Addresses This

**EdgeVerve AI Next is a unified, enterprise-grade platform** that brings together the full lifecycle of agentic AI—from building and orchestrating agents to evaluating, governing, and monitoring them in production. Rather than requiring enterprises to assemble a patchwork of specialized tools, EdgeVerve AI Next provides evaluation as a native capability within the same platform where agents are designed, deployed, and managed.

The platform embeds the full spectrum of evaluation methodologies discussed in this paper—LLM-as-a-judge, agent-as-a-judge, predefined scorers, custom business metrics, and human-in-the-loop workflows—into a single, cohesive evaluation engine. This enables teams to define success criteria once and apply them consistently across development, staging, and production environments, without rewriting evaluation logic or maintaining separate toolchains.

## Key Differentiators

### Evaluation embedded in the orchestration layer:

Unlike standalone evaluation tools that operate outside the agent runtime, EdgeVerve AI Next runs evaluations within the same platform that orchestrates agent workflows. This provides direct access to full execution traces—every tool call, reasoning step, and decision point—without requiring separate instrumentation or data pipelines.

### Continuous evaluation across Dev → Prod:

The platform supports both batch evaluations (running comprehensive test suites against curated datasets) and streaming evaluations (scoring live production interactions in real time). Automated regression checks are integrated into deployment workflows, ensuring that agents meet quality thresholds before they reach production and continue to be monitored once live.

### Governance and compliance as a first-class concern:

For regulated industries—banking, insurance, healthcare—evaluation alone is not sufficient without auditability and policy enforcement. EdgeVerve AI Next provides built-in governance controls, risk management guardrails, and compliance enforcement, ensuring that agent behavior is not only evaluated but also constrained within organizational and regulatory boundaries.

### End-to-end observability with actionable insights:

The platform captures full session traces across multi-step, multi-agent workflows, providing deep visibility into not just what an agent produced but how it arrived at each decision. Drift monitoring and anomaly detection surface degradation proactively, enabling teams to intervene before quality issues reach end users.

### In Practice: Document Processing Agent Evaluation

To illustrate how these capabilities come together in a real-world context, consider a scenario where an enterprise is deploying a document-processing agent to extract structured data from complex, unstructured documents—a workflow that demands high accuracy and strict format compliance.

In such a setup, the evaluation strategy could leverage EdgeVerve AI Next's platform capabilities in several ways. Curated ground-truth datasets would be created using representative production documents, capturing expected extraction outputs for each document type. The platform could then run multiple rounds of automated evaluation: an initial pass using predefined scorers to measure extraction accuracy and structural correctness, followed by LLM-as-a-judge assessments to evaluate free-text field quality and reasoning coherence. When automated scores fall below defined thresholds, the system flags outputs for human review—enabling a feedback loop that continuously refines both the agent and the evaluation criteria.

This iterative, multi-layered approach—ground-truth validation, automated scoring, LLM-based quality checks, and human-in-the-loop verification—embodies the evaluation framework outlined in this paper. Critically, all of these evaluation steps are orchestrated within the same platform used to build and deploy the agent, eliminating the integration overhead that plagues fragmented tooling approaches.

## Strategic Guidance for Enterprises

Based on our experience deploying agentic AI across industries, we recommend that enterprises adopt the following approach to evaluation:

**Start with clear success definitions:** Before selecting tools or metrics, articulate precisely what success means for each agent. Vague goals lead to vague evaluations.

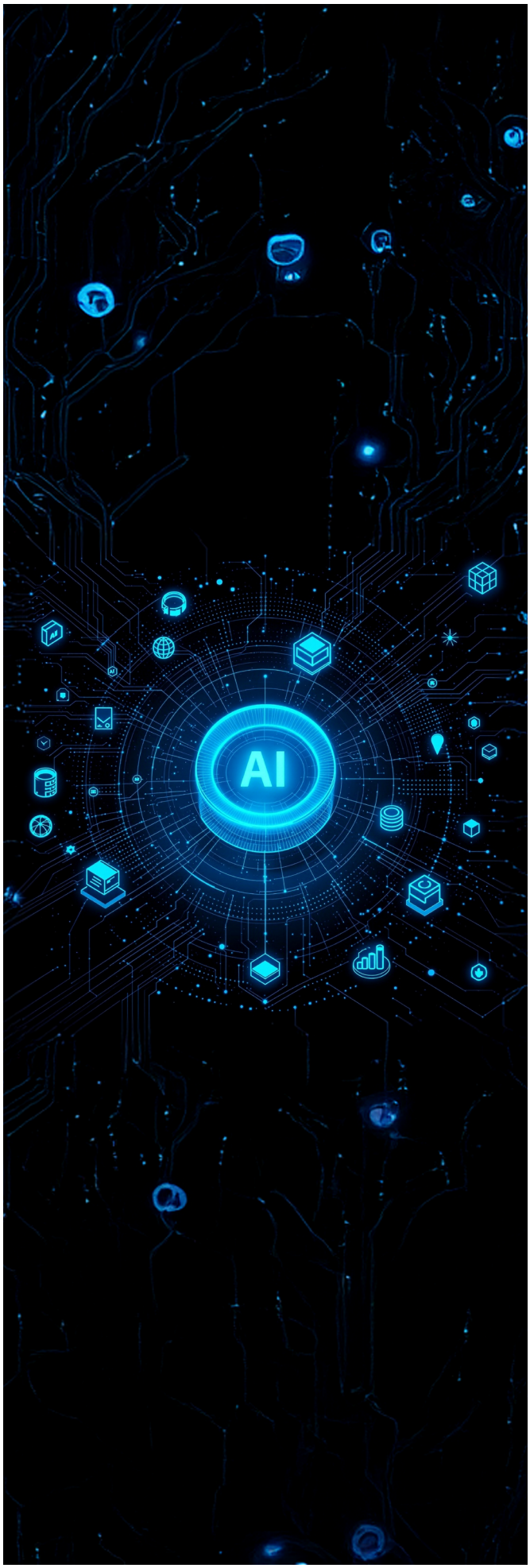
**Evaluate the process, not just the output:** Invest in trace-level evaluation that examines tool selection, reasoning quality, and policy adherence—not just whether the final answer was correct.

**Adopt a layered evaluation methodology:** Combine predefined scorers for speed, LLM-as-a-judge for nuanced assessment, custom metrics for business-specific KPIs, and human review for high-stakes validation.

**Make evaluation continuous, not episodic:** Embed evaluation into CI/CD pipelines and production monitoring. Agents change with every prompt update, model swap, or data shift—evaluation must keep pace.

**Choose a platform, not a point solution:** As the evaluation landscape matures, the greatest risk is fragmentation. A unified platform that integrates orchestration, evaluation, observability, and governance ensures that rigor does not come at the cost of operational complexity.

EdgeVerve AI Next is purpose-built for this approach. By embedding evaluation directly into the agent lifecycle, the platform enables enterprises to scale their AI footprint with confidence—knowing that every agent is continuously assessed, governed, and optimized for quality, safety, and business impact.



## 9. Conclusion

Agentic AI has changed the game. The unit of quality is no longer a single prediction; it is a sequence of choices culminating in outcomes subject to safety, cost, and reliability constraints. Evaluations act as the nerve system that makes such complexity tractable: they encode expectations, detect regressions, and enable safe velocity. But doing evaluations well demands infrastructure, orchestration, governance, and observability—hence the platform imperative.

Platforms like EdgeVerve AI Next operationalize evaluations end-to-end, making rigorous, continuous assessment a built-in property of how agents are developed and run, not an afterthought.

> **Enterprises will trust agents only when their reasoning is as observable as their results. Embedding evaluation directly into the platform is what makes that trust possible—and what ultimately becomes the foundation of AI readiness.**
>
> Sathish Kumar E V, AVP – Senior Director, Product Management, EdgeVerve

**References:**


- Evaluating LLMs/Agents with MLflow | MLflow
- Why Evaluate Agents - Agent Development Kit
- Demystifying evals for AI agents \ Anthropic
- AI Agent Evaluation | DeepEval by Confident AI - The LLM Evaluation Framework
- Agent Evaluation: How to Test and Measure Agentic AI Performance - MachineLearningMastery.com
- LLM-as-a-Judge vs Human Evaluation
- A methodical approach to agent evaluation | Google Cloud Blog

AI Next

edgeverve
An Infosys company

## Ready to explore how this could look like for your enterprise?

Reach out to us at

**contact@edgeverve.com**

AI Next

EdgeVerve AI Next, part of Infosys Topaz, enables the scaling of Applied AI across the enterprise. Built from the ground up to leverage the power of Generative AI, this unified platform bridges silos in people, processes, data, and technology to drive transformation in business operations.
EdgeVerve. Possibilities Unlimited.
https://www.edgeverve.com/ai-next/

edgeverve
An Infosys company

About EdgeVerve
EdgeVerve Systems Limited, a wholly-owned subsidiary of Infosys, is a global leader in developing digital platforms, empowering clients to unlock unlimited possibilities in their digital transformation journey. Our purpose is to inspire enterprises with the power of digital platforms, thereby enabling our clients to innovate on business models, drive game-changing efficiency, amplify human potential, and foster a connected ecosystem. Our comprehensive platform portfolio (EdgeVerve AI Next, AssistEdge, XtractEdge, and TradeEdge) across Automation, Document AI, and Supply Chain helps inspire global enterprises to bridge silos in people, processes, data, & technology, discover & automate processes, digitize & structure unstructured data, and unlock the power of the network by integrating value chain partners. With a deep-rooted entrepreneurial culture, EdgeVerve's innovations are helping global corporations across sectors such as financial services, insurance, retail, consumer and packaged goods, life sciences, manufacturing, telecom, utilities, and more.
EdgeVerve. Possibilities Unlimited.
www.edgeverve.com