

WHITEPAPER

Agentic AI in Production

A QE Governance Framework for Multi-Agent Systems

Table Of

CONTENTS

Executive Summary	1
Introduction	3
Why Multi-Agent Systems Are Categorically Different	4
The Gap: What's Missing in the Market, and What Evaluating Agentic Systems Actually Requires	7
Assurance Dimensions for Agentic Systems: The Five-Layer Framework	10
QE Metrics and KPIs That Matter for Agentic Systems	12
Rollback or Roll-Out: Decision Insights For Go/No-Go Gates	14
Why ValidAlte	15
Conclusion: Making Trust Measurable	16

Executive Summary

Enterprise AI has reinvented itself once again—from the first wave of generative AI (GenAI) systems that answered questions and produced content on the go to the agentic AI systems that can set goals, make decisions, coordinate with other agents, and take real-world actions. Organizations are beginning to realize the value and benefits of multi-agent systems across supply chain orchestration, financial operations, customer service, and IT operations. This shift towards autonomous systems that can function with minimal human intervention creates a governance challenge that most enterprises are not yet equipped to manage.

For instance, when a single model produces inaccurate output, human reviewers can identify and correct it. However, human review cannot eliminate every error that slips through the cracks. And, when a network of autonomous agents executes a wrong task, it may have downstream effects across multiple systems. Agents with task execution privilege can perform actions such as releasing payments with authorization, altering records and statuses, and reconfiguring systems. While incidents can surface only if agents are authorized to perform such actions, many multi-agent systems operate with a strict human-in-the-loop approach, where proposed actions require approval. The argument is that these errors and inaccuracies are difficult to oversee when multiple, independent decisions and actions interact. Moreover, the assurance frameworks built for deterministic software and single-model AI systems cannot be layered on top of agentic systems that act autonomously.

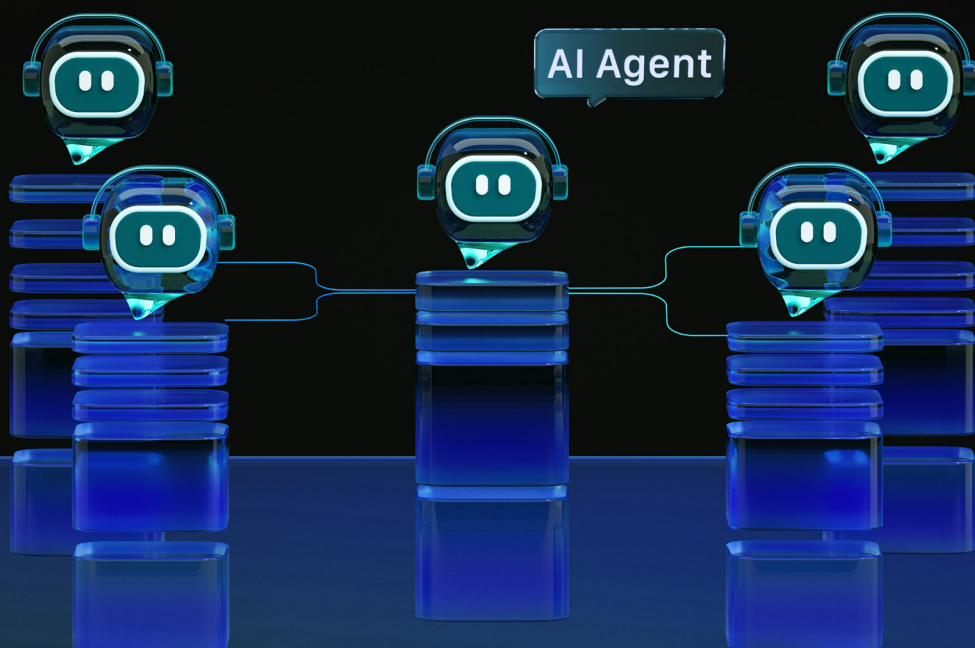
The scale of this gap is widening as it rapidly spreads across enterprise contexts. Beyond anecdotal evidence, Gartner predicts that more than 40% of agentic AI projects will be canceled by the end of 2027. The reasons are not primarily that the underlying technology fails, but rather that costs are escalating, business value is unclear, and risk controls are inadequate.

Separately, a 2026 industry survey found that less than 25% of organizations building AI systems are doing so with the intent of operating independently with minimal human intervention. However, most respondents reportedly insist on keeping humans in the loop precisely because an agentic governance and assurance framework has not yet caught up in terms of capability.

Today, organizations are fervently adopting autonomous multi-agent systems, recognizing the value of improving scalability, modularity, task specialization, and resilience for workflows such as customer support, software engineering, and enterprise automation. Their true value proposition is their ability to orchestrate end-to-end business processes with a level of autonomy that a single-model approach cannot guarantee. While advancements in agent frameworks are emerging alongside the growing relevance of a governance-first framing, what lags is operational Quality Engineering. Organizations are investing heavily in governance, observability, and deployment tooling. However, there is still no standardized testing methodology to help determine and confirm how multi-agent systems perform in real-world scenarios and under complex conditions.

For the multi-agent system development teams and leadership, the challenge is not proving that agents never falter; it's about promptly identifying, containing, and remediating failures before they lead to undesired actions at the enterprise scale. This means QE teams need more than repeatable evaluation methods; they need a structured validation workflow that provides proof that the overall agent workflow is executed as intended.

Your QE teams would already have gathered enough proof points about failure rates. They need a fresh upgrade of their QE discipline that also considers a slew of risks – hallucinations that propagate across the agent value chain, agent-to-agent communication failures due to context loss, inconsistent logic or conflicting plans, workflow failures such as incorrect task delegation and



misconfigured dependencies, and excessive agent privileges. It would go further by defining quantifiable quality indicators such as task completion rates, traceability of agent decisions and underlying logic, consistency over mundane, repeatable executions, and transparency about failure points within the agent network.

At QualiZeal, we do not embrace the 'wait-and-see' approach or claim to have solved the challenge of missing operational QE in these dynamic environments. We approach this with appropriate humility, offering a practical, governance-first point of view: a structured way to think about quality engineering for multi-agent systems, grounded in our experience building AI assurance capabilities through ValidAlte, and informed by the current governance and standards literature, including the NIST AI Risk Management Framework, NIST's Test, Evaluation, Verification, and Validation (TEVV) discipline, ISO/IEC 42001, and the EU AI Act.

Read this whitepaper to understand how QualiZeal proposes a five-layer QE Governance Framework spanning agent validation, interaction validation, tool validation, governance validation, and production monitoring. Key takeaways include:

1. How are multi-agent systems categorically different? Why multi-agent systems need governance?
2. Beyond tall claims, what's missing in the market for evaluating agentic systems? What are the current governance gaps in a multi-agent system?
3. What are the agentic-specific quality metrics that make trust more measurable than aspirational metrics?
4. QualiZeal's five-layer assurance dimensions for multi-agent systems.

Introduction

Why Now: The Scale of Multi-agent Deployment Makes Governance A Real Urgency

Agent explosion is a reality in 2026. According to McKinsey's 2025 State of AI report, nearly a quarter of organizations were scaling agentic AI systems across at least one business function within their enterprise. A recent Gartner insight finds that nearly 80% of enterprises reported that at least one production application embeds AI agents. In high-stakes industries like banking and insurance, the production deployment rates of agentic AI are at 47%, according to S&P Global Market Intelligence and McKinsey.

The healthcare industry also shows a strong uptick in AI agents embedded in its knowledge management, with 14% of organizations scaling agents in that function. Additionally, KPMG's research on AI usage in healthcare shows the technology's biggest impact in IT, with 68% organizations deploying strategies explicitly favoring the use of AI agents across end-to-end decisions. Like every AI conversation that eventually reaches the same tensions and speculations, enterprise AI agents have crossed adoption thresholds in the first quarter of 2026, making governance a real urgency. Because the speed at which organizations move is directly proportional to the opacity of the agents in confirming what they are actually doing. That means the greater the gap, the more it would cost them once something goes wrong in production rather than in controlled environments.

Agent-based Systems are a Production Reality

According to Gartner, 40% of enterprise applications will integrate task-specific agents by the end of 2026, transforming their role from tools that support individual productivity into platforms enabling seamless automation, collaboration, and dynamic workflow orchestration. Integrating agents into enterprise applications is one of the strongest indicators that agent-based systems are moving from the experimental phase to production reality.

Enterprise Digital Workforce will Power Thousands of Agents (Concurrently)

IBM's latest survey, presented at Think 2026, shows that by the end of this year, most large enterprises will have deployed over 1,600 AI agents. With a sharp rise in scalability volume, 70% of executives surveyed reported structural strain and limitations in their current AI governance as a primary bottleneck to the AI transformation. The insights further dived into a staggering trend. While enterprises vigorously pursue ambitious (but isolated) AI projects involving autonomous agents across the enterprise, moving in individual directions can guarantee individual success but will not directly impact the organization. With only 18% of organizations maintaining a current and complete repository of AI, and 68% of executives worried that these initiatives are likely to tank due to a lack of deep integration, the concerns about ungoverned multi-agent systems are hypothetical.

Risks of Agent Sprawl are More Severe than Shadow IT

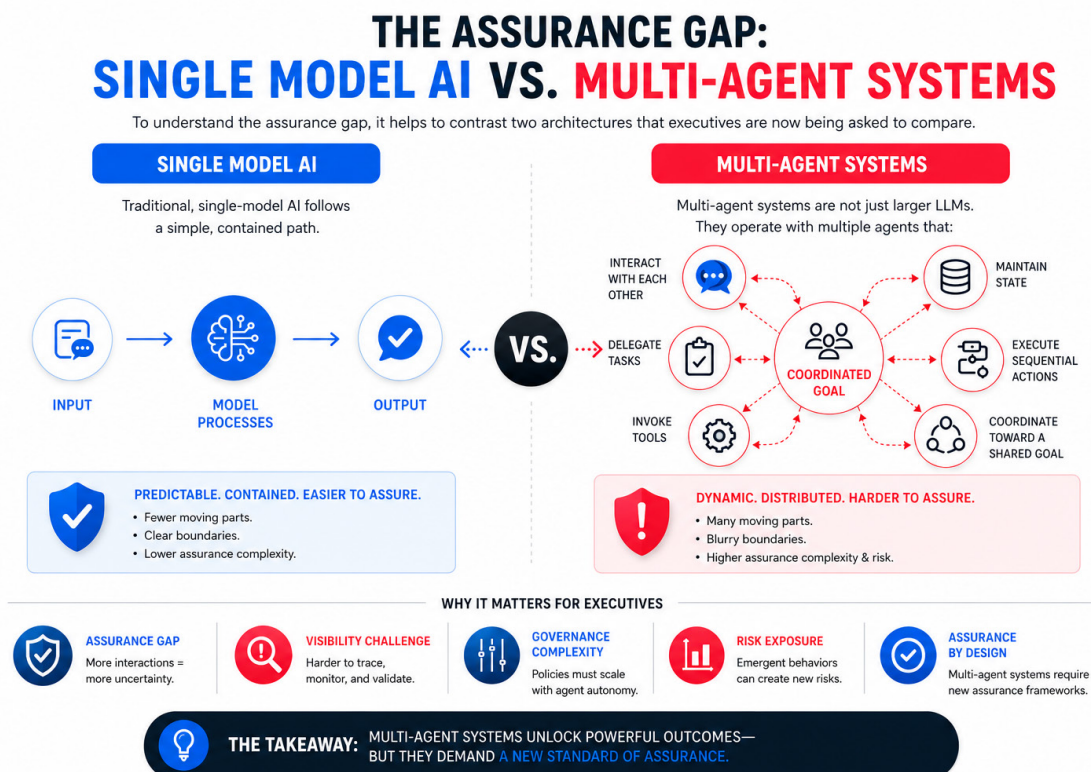
A 2025 study by Enterprise Management Associates found that about 79% of organizations lack formal security policies for their autonomous tools and systems. With agentic systems spun up across business units, departments, and individual employees, it is becoming increasingly difficult for IT teams to keep a close watch. Agent sprawl is more dangerous than shadow IT, as these systems don't just house data but also independently share and interpret it while interacting with production systems. What makes this more concerning is that AI agents have enterprise-wide access and permission, tool access, and API credentials. The financial exposure from ungoverned and shadow AI can reach \$670,000 (IBM study), compared to the breaches involving sanctioned AI tools.

Ungoverned agent sprawl increases vulnerability to a range of risks, such as the spread of misinformation, unauthorized data sharing, and data loss.

Why Multi-Agent Systems Are Categorically Different

To understand the assurance gap, it helps to contrast two architectures that executives are now being asked to compare. Multi-agent systems are not just larger LLMs. Unlike traditional, single-model AI, which follows a simple, contained path: an input is received, a model processes it, and an output is returned, multi-agent systems operate with several agents or sprawl that:

- Interact with each other
- delegate tasks
- invoke tools
- maintain state
- execute sequential actions
- coordinate with each other towards a shared goal.



Additionally, agentic AI replaces single-step actions with a chain of autonomous activities. There is no singular path from input to output—it is a dynamic, branching workflow whose exact route changes from one execution to the next. In simple QE terms, the evaluation is not limited to the model but the entire system.

This structural change introduces several categories of risk that have no equivalent in traditional software testing:



Dynamic and emergent behavior:

Multiple agents interacting can produce collective outcomes that none of them would produce independently. Meaning, they cannot be predicted from testing each agent in isolation.



Agent conflicts:

Agents pursuing overlapping goals may sometimes contradict, block, or reverse one another's work, creating deadlocks or inconsistent results.



Hallucination propagation:

Multiple agents exchange information and make decisions. A confident, plausible, but incorrect output from one agent becomes trusted input to the next, allowing a single error to cascade through an entire workflow.



Goal drift:

An agent may reinterpret a high-level objective in ways that diverge from intent, optimizing for the wrong thing while appearing to make progress. Further, multi-agent workflows involving several sequential steps increase the likelihood that mistakes and misinterpretations will propagate.



Autonomous actions:

The system has permission to take consequential steps and act on decisions, executing transactions, configuration changes, and communications that would have real-world implications before human reviewers would notice them.



Errors attribution Blindspot:

When agents make mistakes, it becomes hard to determine which agent committed the error, whether another agent amplified it, or whether the error emerged from cross-agent interaction.



Uncontrolled tool usage:

Broad access to APIs, databases, and execution environments expands the blast radius of any single mistake or manipulation.



Academic Evidence on Multi-agent System Failures

A 2025 research paper at UC Berkeley published a report on why multi-agent LLM systems (MAS) fail, highlighting the need for systematic identification and analysis of failure patterns. They built the first Multi-Agent System Failure Taxonomy (MAST) through rigorous Grounded Theory-based analysis of a high-quality collection of 150 traces, which were closely reviewed and guided by human annotators and confirmed by high inter-annotator agreement ($\kappa = 0.88$). This method helped detect 14 unique failure modes, clustered into three categories: (i) system design issues, (ii) inter-agent misalignment, and (iii) task verification.

Their analysis revealed a 41% to 86.7% failure rate on 7 state-of-the-art open-source multi-agent systems, with no clear consensus on how to build robust systems. The study also identified three reasons for individual failure modes by frequency: step repetition (15.7%), reasoning-action mismatch (13.2%), and agents being unaware of termination conditions (12.4%). The MAST provides empirically rich evidence for understanding failure dynamics, along with standardized language and a framework for addressing multi-agent risks. In summary, the study also confirmed that enhancements to base model capabilities will not be sufficient to address the full taxonomy, highlighting the need for system-level design and governance, not just model quality.

The Gap:

What's Missing in the Market, and What Evaluating Agentic Systems Actually Requires

Several major enterprises and industry stalwarts have published real, named platforms and frameworks covering agent orchestration, policy enforcement, and observability at enterprise scale. However, most enterprises govern AI with written rules, policy documents, executive system prompts, and constitutional instructions baked into the model. This approach helps define and declare what an agent should and shouldn't perform. This is defined as declarative governance. However, this is likely to fail in agentic systems, as agents are more inclined to follow the rules only when they are easy to decipher. For instance, when optimizing towards a goal, if breaking the rule makes it a more efficient and feasible path, the agents are more likely to bypass it. This is not because the agents are encoded to act maliciously; they are built to pursue the objective. The optimization pressure affects AI agents across thousands of decisions, enabling them to act on goals without performance fatigue, guilt, or social consequences.

Telling an agent what to do is one aspect of damage control, but not complete governance. Below is a list of what is currently missing in the current market when it comes to multi-agent governance:





Gap 1: No Integrated Governance Framework Exists for Multi-Agent Systems Specifically

Every major AI governance framework and risk playbook in use today—NIST AI RMF, ISO/IEC 42001, and the EU AI Act—has legally binding, risk-tiered mandates for enterprises. Because agents engage in action planning, retain memory, and autonomously execute decisions without deep human intervention, these standard frameworks designed for static AI models do not work for autonomous systems. NIST recently launched the AI Agent Standard Initiative, focusing on agent identity, least-privilege authorization, and prompt intrusion defense. To manage compliance across major regulations in the U.S. and the EU, enterprises need a solid unified governance framework that integrates risk protocols with an internal risk taxonomy, audit, and documentation measures that satisfy regulatory requirements, and align agent identity and authorization with emerging NIST AI Agent standards.

What QE assurance must cover: What's clearly missing is an end-to-end evaluation methodology that governs the orchestration layer, not just individual agents. Because no regulatory or standards body has yet defined what that looks like in practice.

Gap 2: No Standardized Evaluation or Testing Methodology for Agent Reliability

Enterprises cannot consistently test, benchmark, or verify whether multi-agent systems are reliable before deployment. Existing evaluation tools can measure task success at the endpoints. However, they do not assess the quality, safety, or compliance with governance requirements of the reasoning path that produced it.

What QE assurance must cover: QE providers with expertise in assuring multi-agent systems provide standardized evaluation protocols that test the path, not just the outcome, while covering tool-call accuracy, inter-agent handoff quality, hallucination propagation across agent chains, and failure mode detection across all 14 MAST categories.

Gap 3: Agent Identity Is Ungoverned at Enterprise Scale

Every agent acting in an enterprise environment uses credentials, holds permissions, and makes authorization decisions in real time. Yet enterprises have no coherent process for treating agents as governed, accountable identities. Agents are routinely given the same generic service-account credentials as traditional software, like the credentials that don't expire, aren't scoped to specific tasks, and leave no decision-level audit trail.

What QE assurance must cover: Enterprises can benefit from testing frameworks that validate agent identity scoping, credential lifecycle management, least-privilege enforcement, and sub-agent delegation chains, both before and after deployment.

Gap 4: Observability Stops at Logs, Not at Decisions

Most observability tools can reveal what an agent did. They cannot correctly answer the governance question that actually matters: why did it decide to do that, and was that decision within sanctioned boundaries? Without decision-level traceability, forensic reconstruction of a multi-agent incident is structurally incomplete, and regulatory audit readiness is impossible.

What QE assurance must fill: QE for multi-agent systems will include decision-level traceability across every agent handoff, capturing not just action logs but the reasoning state and logic, tool selection rationale, data accessed, and delegated authority at each step. These insights will be captured in a regulatory and audit-ready format.



Gap 5: Human Oversight Is Nominally Present but Operationally Superficial

Regulators mandate meaningful human oversight and judgment. What enterprises have built in practice is humans nominally placed in workflows towards the end, when the agents make errors, or something is off the mark. Human reviewers are positioned where they cannot actually intervene—because the pace, volume, and opacity of agent decision-making exceed any human's capacity to review it meaningfully. This translates to accountability transfer with meaningful control.

What QE assurance must fill: Include calibrated human review protocols that define which decisions require human intervention based on risk level, irreversibility, and regulatory exposure. Those insights must be transformed into measurable metrics to confirm that the review is substantive, not ceremonial.

Gap 6: Cascading Failures Across Agent Chains Have No Dedicated Detection Method

In a single-agent system, a failure is bound to one area or action. In a multi-agent system, one agent's error can be the next agent's input. Failures compound, amplify, and produce outcomes that no individual agent's governance controls, anticipates, or could have prevented, because no one governs the interactions; only the individual components are governed.

What QE assurance must fill: System-level testing that evaluates the interaction layer, including agent-to-agent handoffs, shared context integrity, and correlated failure detection across agents using the same base model, and propagation mapping across the full pipeline.

Gap 7: Agent Sprawl Creates Ungoverned Attack Surface That Grows Faster Than Inventory

Enterprises cannot govern what they cannot see. Low-code and no-code tools now allow any employee to spin up an agent in minutes. However, the result is a proliferating, ungoverned agent population that security and governance teams have no visibility into. This enables the attackers to target, impersonate, or manipulate.

What QE assurance must fill: Agent discovery and inventory protocols as a governance prerequisite that have continuous cataloging of every agent in the estate, with ownership, permission scope, risk tier, and lifecycle status, before any evaluation framework can operate effectively.

Gap 8: Regulatory Compliance Cannot Be Demonstrated Without Evidence Architecture

Regulators demand audit-ready evidence of AI governance. Enterprises cannot produce it because the evidence doesn't exist, and not because of bad intent. The inability to produce results stems mainly from the fact that the multi-agent systems were deployed before the evidence-generation infrastructure was designed and implemented. Governance retrofitted after deployment is structurally incomplete and vague.

What QE assurance must fill: Evidence generation built into the testing lifecycle must include structured compliance records, decision lineage, test artifacts, and audit trails generated systematically as byproducts of governance-integrated evaluation. They should not be assembled retrospectively under audit pressure.

Assurance Dimension for Agentic Systems: The Five-Layer Framework

LAYER	ASSURANCE FOCUS	VALIDATES
 Agent Validation	Does each agent do what it is supposed to do?	Goal adherence instruction compliance, role boundaries
 Interaction Validation	Do agents work together correctly?	Agent-to-agent communication, context transfer, workflow integrity
 Tool Validation	Are external actions safe and authorized?	Tool invocation, API safety, external integrations
 Governance Validation	Can humans oversee and account for the system?	Human-in-the-loop, auditability, explainability approval controls
 Production Monitoring	Does the system stay trustworthy over time?	Drift detection, behavioral change, runtime assurance

Layer 1:

Agent Validation:

The foundation is assurance that each agent behaves as expected, remains faithful to its assigned goal, complies with its instructions and constraints, and respects the boundaries of its defined role. The agent quietly expanding its own scope is a governance failure, even if every action it takes appears locally reasonable.

Layer 2:

Interaction Validation:

Multi-agents' intrinsic value comes from collaboration, and collaboration is where emergent risk concentrates. This layer validates whether the interaction between agents is reliable, that context and intent transfer from one agent to the next without distortion, and that the overall workflow retains its integrity when work passes across agent boundaries. This is where hallucination propagation and agent conflict are caught, making failures invisible when agents are tested one at a time.

Layer 3:

Tool Validation:

Tools are where an agentic system touches the real world, and therefore where actions become consequential. This layer ensures that tool invocations are appropriate and authorized, that external API calls are safe and scoped to the least privilege, and that integrations cannot be turned into a vector for unintended or malicious actions.

Layer 4:

Governance Validation:

Autonomy without accountability is unacceptable in an enterprise setting. This layer confirms that human and organizational controls genuinely work: that human-in-the-loop checkpoints trigger when they should, that every consequential action leaves a complete and tamper-evident audit trail, that decisions can be explained to a reviewer or regulator, and that approval controls cannot be bypassed.

Layer 5:

Production Monitoring:

Agentic systems learn, adapt, and operate in changing environments, so assurance cannot stop at launch. This layer provides continuous monitoring for behavioral drift, detection of changes that may signal degradation or emerging risk, and runtime assurance that the system continues to operate within its approved boundaries. Validation becomes a lifecycle activity rather than a one-time gate.

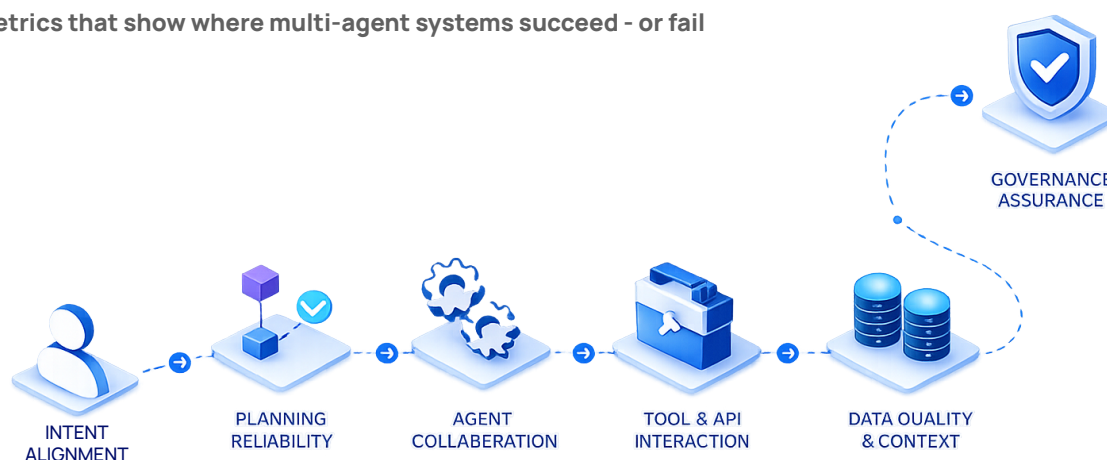


QE Metrics and KPIs That Matter for Agentic Systems

Governance frameworks earn credibility only when they are measurable. The metrics below are organized around the points in a multi-agent system where failure actually originates. These are directly informed by the evaluation dimensions of QualiZeal's ValidAlte™, the enterprise AI assurance systems that already test multi-agent systems in production engagements.

Governance That's Measurable,

Metrics that show where multi-agent systems succeed - or fail



Retrieval and grounding (where applicable – RAG-augmented agents)

- Evaluation metrics for search engines and recommendation systems like Precision@k, Recall, Mean Reciprocal Rank (MRR), and Zero-match rate helps determine whether the agent retrieves the right context, completely and on time.
- Faithfulness/groundedness and citation accuracy helps understand whether the agent's output is actually supported by what it retrieved or is it inventing support for a claim.



Agent behavior and orchestration

- Plan completeness and role fidelity to understand if the agent remains contained in its assigned role and whether it can complete its part of the plan.
- Handoff integrity to understand if the context and intent survive the transition from one agent to the next without loss or distortion and misinterpretation.
- Deadlock rate to assess how often do agents pursuing overlapping goals can stall or block one another.

Tool use and external action

- Invocation accuracy to observe if the agent selects the correct tool for the task.
- Parameter integrity to verify of the tool is valid and safe for use.
- Side-effect incident rate to check how often does a tool call produce an unintended real-world consequence

Trust and safety

- Hallucination rate and contradiction rate to note the frequency at which the system produces fabricated or self-contradictory claims, and critically, how often does that error get passed downstream to another agent rather than caught at the source (hallucination propagation).
- Bias test pass rate and paired-prompt disparity to verify whether the system treat comparable users or cases consistently.
- Refusal quality score and false-positive rate to understand if the safety guardrails blocking genuinely unsafe requests without also blocking valid ones.

System-level reliability

- Agent Interaction Failure Rate to understand how often communication or coordination between agents breaks down end-to-end.
- Workflow Reliability Score to check end-to-end success rate of multi-step action chains under real operating conditions, not just under controlled test scenarios.
- Human Override Frequency for how often a human has to intervene; tracked over time, this is one of the clearest available signals of whether a system is ready for greater autonomy or not.
- Traceability Score and Governance Compliance Score to see if there is a complete, tamper-evident audit trail for every consequential action, and whether approval/policy controls are actually enforced in practice rather than only on paper.

Rollback or Roll-Out: Decision Insights for Go/No-Go Gates

One of the hardest moments in any agentic AI deployment for the enterprise IT leaders and AI boards is not the build. In fact, it is the decision to ship, hold, or roll back. Three structural decisions repeatedly determine whether that go/no-go gate is meaningful:

Decide What “Good Enough” Means FOR Multi-Agent Systems:

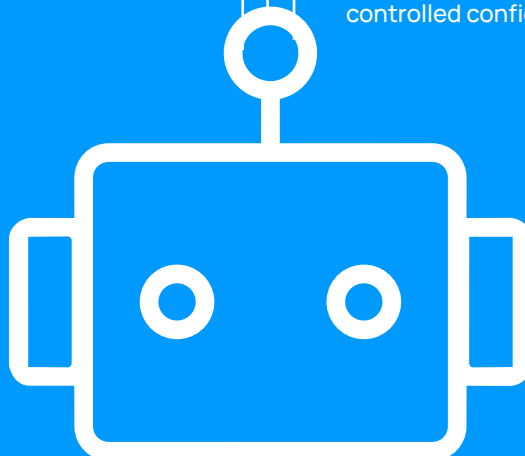
Agentic systems do not guarantee 'no downsides. In fact, they involve trade-offs with no universally correct setting. The onus of defining the agent objectives, goals, and decisions lies with the agentic systems development teams to ensure they are not deliberately set against business priorities. However, agent evaluation and QA teams must look at aspects like accuracy versus creativity (higher faithfulness typically means less novelty in output). Latency versus reasoning depth (lower latency is cheaper and faster but allows less reasoning). Safety versus usability (tighter guardrails reduce risk but can block valid requests). Explainability versus performance (more logging and interpretability add overhead). Personalization versus privacy (better-tailored experiences can increase the risk of data exposure). None of these trade-offs is wrong to make. However, making them implicitly, under deadline pressure, or at the spur of the moment, is how avoidable incidents happen. The trade-off position should be a deliberate, pre-agreed business decision and agent logic, not default settings that can get bypassed in production scenarios.

Use Multi-Dimensional Trust Scores, Not A Single Pass/Fail Gate:

A multi-dimensional trust score is not a 'single percentage' or 'pass/fail' indicator. Instead, it aggregates assurance results across multiple dimensions including agent behavior, orchestration quality, tool safety, governance controls, trust and safety metrics, and production monitoring. Each dimension is evaluated independently and contributes to an overall readiness profile, allowing decision-makers to determine whether a system is production-ready, requires targeted remediation, or should be rolled back for specific capabilities rather than the entire deployment.

Treat Human Oversight Frequency as a Leading Indicator, Not a Failure Signal:

A rising rate of human review in a specific agent workflow is not necessarily evidence that the system is broken. It helps understand exactly where autonomy has not yet been earned. When tracked over time and by failure category, the frequency of human review as a metric is one of the most reliable signals. This will help QE teams report whether the system is ready for autonomy and permission expansions, should it stay within its current scope, or rolled back to a more controlled configuration.



Why ValidAlte

QualiZeal's ValidAlte was built around a simple premise: assurance must be embedded directly in its design, and no AI system should go into production without validation. It is positioned as a complete solution that combines four elements—Platform, Process, People, and Compliance/Governance, rather than a point tool that addresses only one layer of the problem this paper has described.

Platform

ValidAlte's core workflow, Trust, Risk, Metric, Test, Evidence, Observe – is QualiZeal's proprietary operating model for AI assurance. It operationalizes the five-layer governance framework directly by running structured AI evaluations across dimensions including RAG retrieval and generation, agent behavior, tool calling, conversational quality, bias, and hallucination. Rather than producing a simple pass/fail output, the platform generates explainable evidence and easily integrates with enterprise evaluation, and observability tools already present in existing enterprise AI stacks.

Process

Delivery runs through a structured pod model comprising of AI risk and governance assessors, GenAI evaluation engineers and leads, domain experts, automation engineers, and EvalOps engineers working together. The pod model follows a structured path that steps away from treating AI assurance as a generic testing task bolted onto an existing QE team.

People

The pod model deliberately embeds domain experts in the evaluation loop, consistent with the lesson that evaluations grounded in real domain data outperform generic, one-size-fits-all metrics for any system handling specialized or regulated workflows.

Compliance and Governance

While Trust, Risk, Metric, Test, Evidence, Observe serves as ValidAlte's operational assurance lifecycle, the platform is aligned with established governance frameworks including the NIST AI Risk Management Framework (Govern, Map, Measure, Manage), ISO/IEC 42001, and applicable regulatory requirements such as the EU AI Act. ValidAlte generates trust metrics, authorization trails, continuous monitoring records (ref. section 4 & 5 of this paper), and audit-ready evidence that support these frameworks without replacing them.

This is dogfooded internally before it is sold externally: ValidAlte's evaluation methodology is production-proven on QMentisAI, QualiZeal's own GenAI platform for quality lifecycle management, and informs NexaAI, QualiZeal's enterprise GenAI development services – meaning the assurance approach described in this paper has been tested against QualiZeal's own production systems, not only client engagements.

The result, demonstrated in real deployments: organizations working with ValidAlte have seen up to 40% in operational cost savings, alongside meaningful gains in quality and speed to market across LLM chatbot validation, RAG pipeline validation, and agentic development validation engagements.

Conclusion: Making Trust Measurable

Agentic AI is no longer an enterprise curiosity; it is moving into production across the critical value, faster than the assurance practices meant to govern it. With that shift comes a change in what QE must achieve—moving reactive testing approaches suitable for testing traditional systems software or single model AI systems to assuring autonomous agent ecosystems that take actions while reducing the cost of unexamined failure of wrong agent actions and collective hallucinations.

If Gartner's forecast—40% of agentic AI projects will be canceled by 2027, is read correctly, the priority is not about AI capability. It highlights the strong implications from the absence of demonstrable agent assurance and governance frameworks.

At QualiZeal, we don't claim to have solved agentic (multi-agent) AI assurance, and this paper has tried throughout to be precise about the difference between what the market has published and what remains genuinely open. We propose a practical, governance-first QE framework involving five layers, a measurable set of trust metrics, and direct alignment with NIST AI RMF, TEVV, ISO/IEC 42001, and the EU AI Act.

The overarching idea is to advance enterprises to move their multi-agent AI systems into production safely and responsibly, with relevant assurance capabilities, evidenced in production engagements, to support that journey.

The bottom line: trust must be made measurable, not just in end results, but in numbers that enterprise board can review, evidence that external auditor can examine, and a score that changes a go/no-go decision. This is an emerging discipline, and we offer this framework as a credible, defensible starting point alongside our clients, the standards community, and the broader field as it matures.

Soon, enterprises that move first would be those with a structure, measurement, and governance for their agentic systems and can clearly demonstrate what trustworthy autonomous AI looks like.

About QualiZeal

QualiZeal is a leading global AI-native Quality Engineering and AI Assurance partner delivering business-critical assurance for enterprise IT and AI. Founded in 2021, QualiZeal serves 70+ global enterprise clients, including Fortune 500 organizations, and sustains a CNPS of 80+. Its delivery capability is anchored by a 1,000+ person AI-ready workforce spanning Quality Engineering, Digital Engineering, and Emerging Technology Testing.

QualiZeal's approach is automation-first, platform-powered, and AI-native, built to deliver Trust First AI engineering alongside AI-led Quality Engineering. This is operationalized through an AI CoE-led innovation and R&D function that has developed an integrated platform ecosystem that addresses quality through two disciplines: AI-powered QE and QE for AI.

QMentisAI is QualiZeal's homegrown IP-led innovation platform and patent-pending innovation, enabling fully agentic Quality Engineering across the Quality Management Lifecycle, automating and orchestrating intelligent quality workflows. ValidAlte™ is an enterprise AI assurance platform for evaluating trust, safety, governance, and compliance readiness. It brings together leading AI governance frameworks, regulations, standards, and assurance practices—including NIST AI RMF, the EU AI Act, ISO/IEC 42001, and TEVV—within a unified evaluation framework, enabling enterprises to assess and evidence AI trustworthiness across the AI lifecycle. NexaAI QualiZeal's latest service line for enterprises to move from AI strategy to development and operationalization with decision-packs, from advisory, consulting, and engineering services to ship AI projects with proof, not promises.

QualiZeal has been recognized by Everest Group, Gartner, Forrester, Frost & Sullivan, and NelsonHall. Its accolades include the 2026 Gold Stevie® Award for Technology Excellence in AI-Powered Quality Engineering, the 2026 Globee Awards' Artificial Intelligence Achievement of the Year, inclusion on the 2026 Inc. 5000 list of America's fastest-growing private companies, and a ranking on the Deloitte Technology Fast 500™.

Appendix

<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>

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

<https://www.mckinsey.com/featured-insights/week-in-charts/agentic-ai-advances>

<https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2026/06/intelligent-healthcare-report.pdf>

<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>

<https://www.ibm.com/think/news/think-2026-ai-recap>

<https://www.ibm.com/thought-leadership/institute-business-value/us-en/report/ai-orchestration-layer>

<https://www.ibm.com/thought-leadership/institute-business-value/us-en/report/enterprise-2030/>

<https://councils.forbes.com/blog/agentic-ai-sprawl-audit-approve-kill-autonomous-workflows-before-they-multiply>

<https://www.pixee.ai/blog/agentic-ai-governance-gap-strategic-framework-2026>

<https://arxiv.org/pdf/2503.13657>

<https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>

For more information on how QualiZeal can elevate QE
for your enterprise AI systems, contact us at
info@qualizeal.com



<https://qualizeal.com>