Read Time: 20 Mins

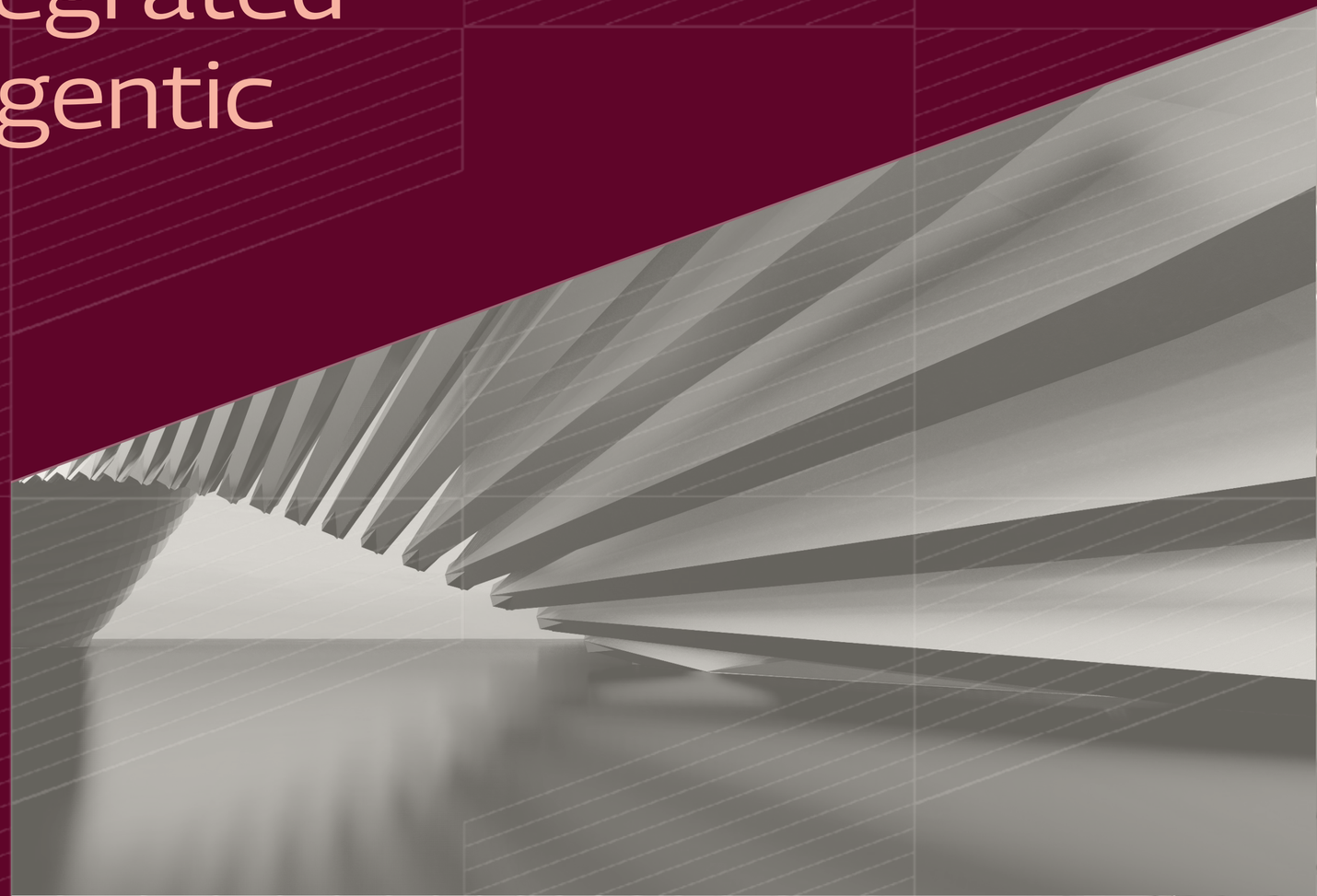
Scale at Speed™

THIRD EYE ADVISORY™

TECH mahindra

Whitepaper

Architecting the Integrated Quality Fabric for Agentic Systems



Anjali Chhabra

Global Head - Practice & Strategic Pursuits,
Agentic Development & Modernization Services,
Tech Mahindra

Pat Sullivan

Expert Advisor,
Third Eye Advisory

Content

Introduction

Page 4

From Standard Test Cases to Trust Contracts

Page 6

The Architecture of the Integrated Quality Fabric

Page 7

The Operating Method

Page 9

The Universe - Every Type of Agent, Every Type of Agentic Workflow, Every Strategy

Page 12

From Unit to Universe - The NAU-A Commercial Model

Page 14

Proof in Practice and Way Forward

Page 16

Conclusion

Page 18

About the Authors

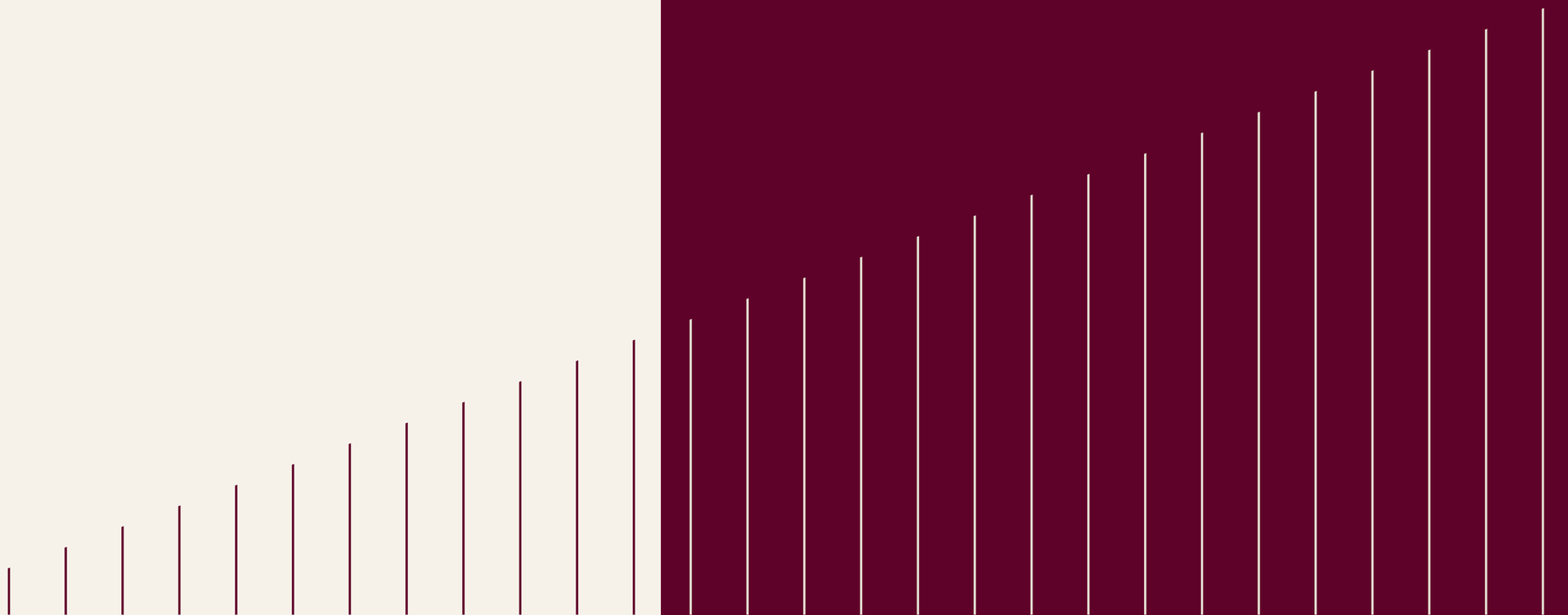
Page 19

References

Page 20

Executive Summary

Agentic systems do not behave like applications: they decide, reason, retrieve, call other agents, learn, and adapt, sometimes producing unreliable or inconsistent outputs over time. Every premise behind traditional quality assurance collapses the moment autonomy enters the enterprise stack. The question for QE teams has shifted from confirming a build passed testing to establishing whether an agent can be trusted to operate and learn at a production scale without human intervention. Tech Mahindra's answer is an integrated quality fabric: a platform spanning unit to universe.



Introduction

Agentic AI is moving from experimentation to enterprise-wide deployment, and the risks are moving with it. Unlike traditional software, agentic systems decide, reason, retrieve, call other agents, and adapt in real time and (occasionally) get it wrong in ways no test script anticipated. That single shift breaks assumptions quality functions have relied on for decades.

Traditional QA was built for deterministic systems: fixed inputs, repeatable outputs, and a clear line between pass and fail. Agentic systems offer none of that; the same prompt can return a different answer tomorrow because context and learning are constantly changing, and quality can degrade silently in production. Most enterprises have responded to the changing nature of quality assurance by patching it: a guardrail here, a policy check there, an evaluation routine bolted onto an existing pipeline. The approach breaks once agents scale into production, across multiple models, tools, and workflows.

An integrated quality fabric offers a different model: treating quality as a continuous, woven layer across the agent lifecycle, not a pre-launch checkpoint. The following sections map where traditional QA collapses, detail how the fabric is architected and operated, and introduce the NAU-A model for predictable, scalable agentic quality assurance.

Where Traditional QA Breaks - Four Collapse Points

Collapse Point	Why traditional QA fails	What the Fabric does
Non-deterministic outputs	Same prompt → different response. Pass/fail is subjective and non-binary.	Behavior envelopes, golden prompts, adversarial sets, similarity scoring.
Multi-agent workflows	Each agent is green; the workflow fails to deliver the outcome. Different defects emerge from the same interactions.	Workflow-level intelligence, multi-agent orchestration simulation, reasoning logic, and decision-trace validation.
Autonomy, reasoning, and decision-making	Agents reason, decide, and take real-world actions. Wrong actions cascade faster than we can imagine.	Enables guardrails, safe-action boundaries and thresholds, confidence scoring, and human-in-the-loop routing.
Drift, hallucination, and leakage	No visible failure; still quality degrades silently in production.	Continuous evaluation, drift detection, tracing and observability, and canary release gates.

We cannot govern the universe of agentic workflow behaviour by bringing together point test cases at the unit. We must weave an integrated quality fabric.

From Standard Test Cases to Trust Contracts in the Agentic Era

Quality, in an agentic era, is no longer a collection of test cases run manually or automated before release. It must be treated as a trust contract that the enterprise signs with regulators, customers, and its own employees. The trust contract ensures that every agent in the system behaves correctly, safely, ethically, and predictably and takes explainable actions. Tech Mahindra's quality engineering practice frames that contract across five dimensions.

The Five-Dimensional Quality Fabric for Agentic Systems

Five independent quality axes woven into one enterprise trust model

Figure 1: The Five-Dimensional Quality Fabric: Each Axis Is Independent, But Trust Requires All Five.

The Architecture of the Integrated Quality Fabric

Embedded, Not Added-On

Most enterprises today treat AI quality as a series of add-ons—a guardrail here, a policy there, an evaluation routine inserted, a telemetry display somewhere else. That approach is acceptable for small POCs or pilots but fails as soon as agents scale beyond a single pilot. Our integrated quality fabric architecture is consciously the opposite: A woven six-layer fabric where every layer is embedded in agentic pipelines, every layer talks to other layers, and our solution accelerators, along with industry-proven tools, thread these layers together.

The Integrated Quality Fabric — Layered Architecture

Embedded quality controls across every layer, supported by a parallel TechM IP rail

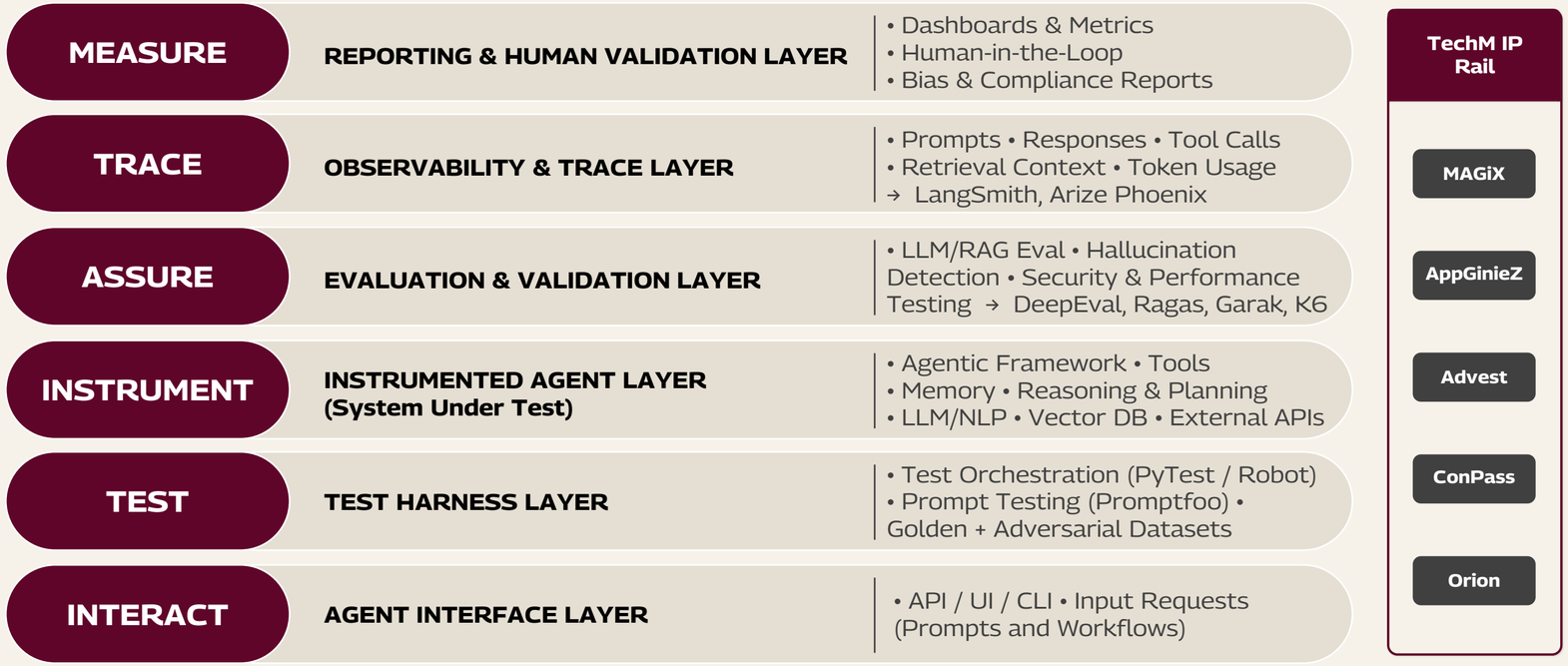


Figure 2: The integrated quality fabric architecture, with Tech Mahindra's IP rail running parallel through every layer.

The Accelerators - What Value Each Accelerator Unlocks

The most important architectural decision: Tooling is composable, not monolithic. We consciously stitched Promptfoo, DeepEval, Ragas, Garak, LangSmith, Arize Phoenix, K6/Locust/JMeter, and Guardrails AI into MAGiX/AppGinieZ. So, enterprises get best-in-class capability per layer, without being locked into any single industry tool or partner's interpretation of 'AI quality.'

TechM IP	What it does	Where it lives
MAGiX Framework	Test orchestration, NLP-driven test design and authoring, low-code functional, and API automation harness for agents and agentic systems.	Test harness and agent interface layers.
AppGinieZ	Agentic and GenAI QE platform—assisted and/or autonomous test generation, RAG validation, multi-agent test design and execution.	Instrumented agent layer and evaluation layer.
Advest	Adversarial test solution integrated with AppGinieZ that verifies LLMs under malicious inputs, bias probes, and jailbreaks.	Evaluation and validation layer.
ConPass	Continuous performance assurance: AI-driven performance engineering across JMeter, K6, and Locust for agentic workflows.	Performance / non-functional.

The Operating Method

Full Lifecycle Coverage: From Inception Through Production

The integrated quality fabric works as a disciplined lifecycle that runs continuously around every agent in production.

Each phase has a specific set of activities, test types, and a specific distribution of effort to construct delivery economics (see NAU-A).

360° Lifecycle Coverage — From Planning to Production Monitoring

One NAU-A distributes 100 hours across seven continuous lifecycle phases

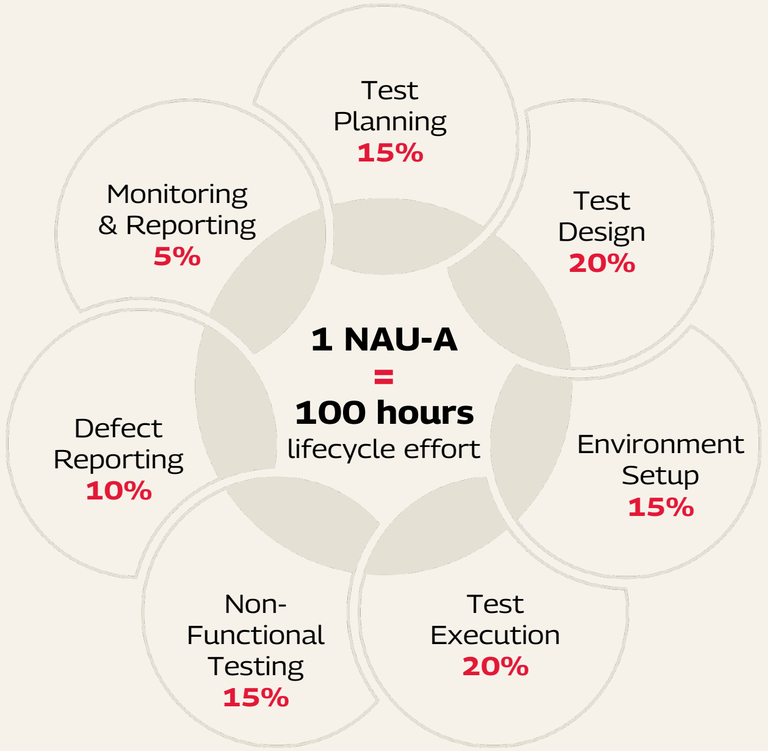


Figure 3: The 360° lifecycle wheel. Effort distribution by phase; 1 NAU-A = 100 hours of integrated lifecycle effort.

Phase-Wise Activities and Effort Distribution

Phase	Key Activities	Test Types	Effort
1 **Strategy and Planning**	Risk identification. • Evaluation of KPIs and metrics. • Dataset selection (golden + edge + adversarial). • Tooling strategy.	Planning, LLM quality, hallucination, and RAG.	15%
2 **Test Design**	Prompt repository. • Evaluation scenarios. • Golden datasets. • Thresholds. • Adversarial and edge cases and alternate flows.	Prompt robustness, adversarial, and integration.	20%
3 **Environment Setup**	Agentic frameworks (LangGraph / CrewAI). • APIs and integrations (MCP servers, A2A). • Vector DBs • Tracing and evaluation pipelines.	Multi-agent observability.	15%
4 **Test Execution**	Prompts/test cases. • Evaluation pipelines. • Trace capture. • Repeated runs. • Reasoning and decision validation.	Functional, LLM quality, RAG, and reasoning.	20%
5 **Non-Functional Testing**	Load. • Stress. • Prompt injection. • Data-leakage • Variance. • Toxicity, policy, and guardrail validation.	Performance, security, reliability, and safety.	15%
6 **Defect Reporting**	Hallucinations. • Reasoning failures. • Decision issues. • Prompt failures. • Defect root cause (model vs prompt vs data).	Defect submission and analysis.	10%
7 **Monitoring and Reporting**	Evaluation dashboards. • Human review, acceptance/rejection/override rates. • Production behavior. • Bias, fairness, and accuracy.	Observability, HITL, and safety.	5%

Beyond Load Testing

Standard load testing tools are built for stateless web requests and standard applications. Agentic systems aren't standard applications or web requests. They stream, reason, decide, retry, summarise context, exhaust token windows, and drift, hallucinate, and degrade silently. ConPass is Tech Mahindra's seven-step continuous performance engineering framework, purpose-built for agents and agentic workloads, with soak and endurance tests (8-24+ hours). ConPass uncovers the failure modes and errors a point-in-time load test can never catch—issues like memory leaks, token explosion, context-window exhaustion, and silent latency creep and drift.

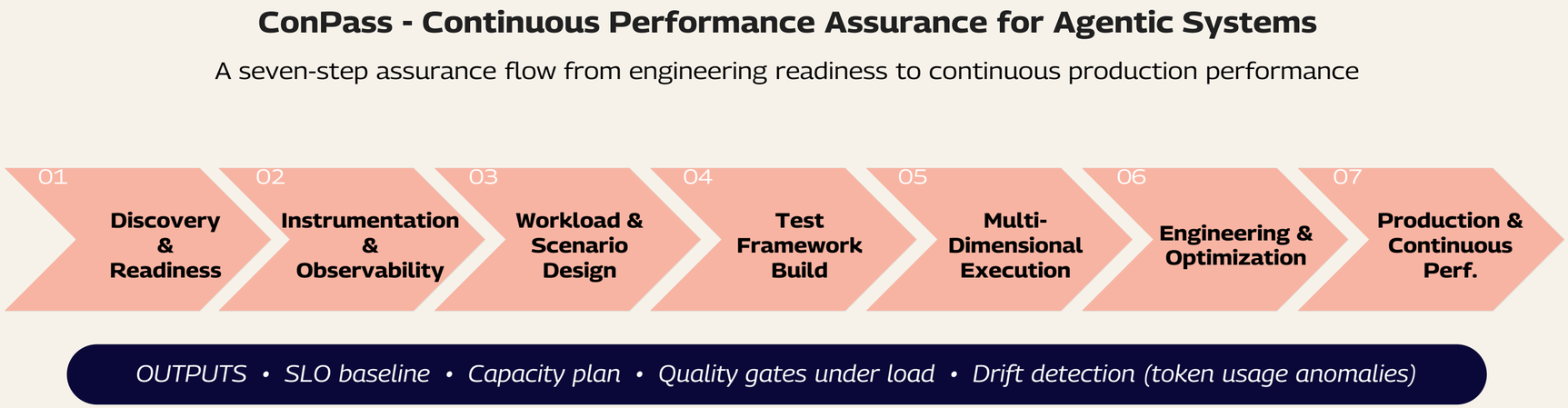


Figure 4: ConPass: seven-step continuous performance engineering, instrumented with canary gates and drift detection.

The Universe - Every Type of Agent, Every Type of Agentic Workflow, Every Strategy

One Fabric, Many Agent Topologies

Enterprises rarely deploy one type of agent. These deployed agents are also based on many different AI platforms. A typical AI estate now includes single-task assistants, single-task autonomous agents, multi-agent workflows, RAG-driven knowledge agents, agents using underlying tools calling APIs and MCP servers, autonomous decisioning and action agents, agentic workflows with human-in-the-loop or human-on-the-loop, and domain-specific agents for support or IT agents (developer and tester agents) themselves. The fabric must test all of them and be adaptable to all the AI platforms on which they are based.

The Agent Universe -TechM Tests Every Class of AI Agent

Six agent classes, one consistent quality fabric

1

FOCUS

Single Agent

Task-focused

Functional • Behavioural • Safety

2

ORCHESTRATE

Multi-Agent

Workflow-based

Functional • Behavioural • Safety

3

GROUND

RAG-Based

Enterprise data / knowledge base

Functional • Behavioural • Safety

4

CONNECT

Tool Integration

MCP / API communication

Functional • Behavioural • Safety

5

DECIDE

Autonomous/Semi-Autonomous

Multi-step reasoning & decisioning

Functional • Behavioural • Safety

6

SPECIALIZE

Domain-Specific

Customer support / QA-Dev specific

Functional • Behavioural • Safety

Every agent class is exercised across functional, behavioural, safety, performance, and observability axes

Figure 5: Six Classes of AI Agents, All Covered By A Single Quality Fabric.

The Strategic Stance for the Integrated Quality Fabric for Agentic Systems-Model-agnostic and Platform-Specific

The single biggest error in enterprise AI today is choosing one platform: pure model-agnostic testing (portable but blind to platform-specific risks) versus pure platform-specific testing (rigorous but locking the enterprise in). Tech Mahindra's stance is conscious: do both.

	Model-Agnostic Approach	Platform-Specific Approach
What it does	Evaluates agents independent of the underlying LLM, sometimes even using the LLM as a judge model, or using common metrics via external evaluation frameworks (DeepEval, Ragas).	Validates features, constraints, and behaviors unique to a specific AI platform—using native platform tools, models, policies, and outputs.
Strategic benefit	Vendor independence and neutrality. Reduces vendor lock-in risk. Future-proof QA governance. Ideal for multi-vendor hybrid LLM estates.	This approach ensures production readiness. Validates security, compliance, and cost controls specific to the platform. Required for many domain native agentic platforms.
When to use	Cross-team evaluation, vendor selection, model selection and efficacy comparison, and cost analysis for switching models.	Highly regulated industries, high-risk platform-dependent workflows, SLA-bound and platform-specific deployments, and audit-mandated environments.

The hybrid approach balances flexibility with platform assurance, avoiding the trade-offs of relying on either approach alone.

From Unit to Universe—The NAU-A Commercial Model

Quality That Is Scalable With Cost Predictability

Most proposals for testing agents or agentic systems can fall apart in decision-making discussions. McKinsey's 2025 State of AI survey found that only a third of enterprises have scaled AI beyond pilots, and just 39% report measurable EBIT impact [2]. The risks, costs, and controls for agentic systems are difficult to understand, opaque, and non-deterministic in scope, and teams cannot clearly predict whether they are investing in 'three engineers for six months' or 'trust in fifty agents for three years.'

Tech Mahindra has solved this with the normalized agent unit (NAU-A), an explainable, calibrated unit that captures the standardized effort to verify, validate, and assure one agent across its functional, behavioral, and operational dimensions. Tech Mahindra has developed a scoring system that addresses business value, complexity, risk, and agent performance to better manage risks and quality.

Calibration Dimensions (V1, V2, V3)

V1: BUSINESS CRITICALITY

LOW	Functional and end-user acceptance, user or business journey coverage.
MEDIUM	User experience, SLA, and response time sensitivity.
HIGH	Business and financial risk of failure, compliance, and regulatory assurance.

V2: WORKFLOW COMPLEXITY

LOW	System interdependence, few integrations.
MEDIUM	Functional orchestration with operational controls, guardrails, and complex integrations.
HIGH	Context assembly across multi-agent workflows, highly complex workflow with HTIL touchpoints, and escalation paths.

V3: AGENT MATURITY

LOW	Model validation maturity.
MEDIUM	Prompt engineering maturity and operational safety.
HIGH	Model sophistication and deployment pattern.

One NAU-A maps to 100 hours of integrated lifecycle effort—split across the seven phases (15/20/15/20/15/10/5%).

This becomes the lowest practical and explainable unit of agentic quality, and it scales almost linearly into agentic workflow delivery economics. A multi-agent ITOps platform with validation and verification effort calibrated against NAU-A becomes a transparent cost-and-quality model, not a black box.

The 5-Year Cost Curve: Why This Model Compounds

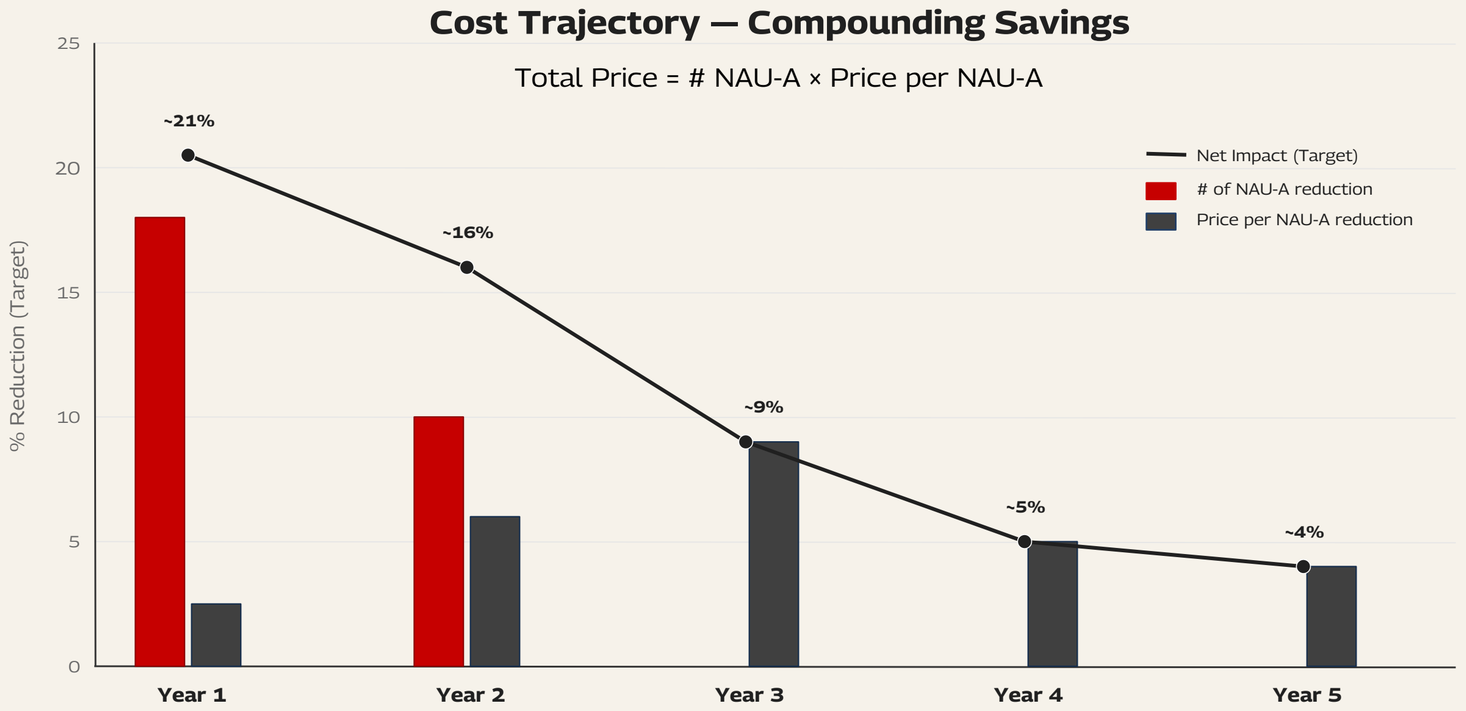


Figure 7: Two compounding levers: reducing # of NAU-A (foundational, Y1-Y2) and reducing price per NAU-A (Y3-Y5 via agent + human augmentation).

The economic story is two-phased. In years 1- 2, deployment of foundational levers of improved prompt engineering, golden dataset curation, agentic frameworks, and toolsets sets up and automates NFT, driving cost reduction in the number of NAU-A consumed per agent. From year 3 onward, the model flips as agent testers begin to augment human testers for agentic workflows, driving down price per NAU-A annually. Cumulatively, teams see a compounding effort-saving curve that traditional staffing models cannot match.

Quality finally becomes a measurable asset on the balance sheet, not a recurring and ever-increasing cost on the P&L. Also, adding every new agentic workflow doesn't have a linearly increasing cost impact.

Proof in Practice and Way Forward

Where The Fabric Has Already Been Delivered

The integrated quality fabric cannot be perceived as a roadmap. It has already been delivering results in production across global enterprises, spanning autonomous ITOps, testing for a large social-platform's agentic AI-based system, regulated railroad operations, mobile-device automation, and large-scale QE estates with a combination of standard applications and agentic workflows. The proof points below illustrate the same fabric, calibrated to different industries, different agent topologies, and regulatory complexities.

Proven Outcomes Across Agentic AI Deployments

Large Telecom: Autonomous ITOps

- Business risk: cascading outages, low operator trust in autonomous remediation
- Business value delivered: 30–35% fewer false auto-remediations; governed autonomy model IT could rely on

Large Social Networking Platform

- Business risk: bias, drift, and ethics failures reaching users at scale
- Business value delivered: 40–45% better defect detection; drift caught before production, protecting against reputational and compliance exposure

Large North American Railroad

- Business risk: slow, effort-heavy QE tying up capacity across customer, payroll, and waybill programs
- Business value delivered: ~30% faster test cycles, ~50% less manual effort; 12-person team down to a 5-day cycle

Leading Technology Company: Mobile Device Testing

- Business risk: device-control complexity slowing time-to-market for automated workflows

Way Forward: Seven Moves for Setting up an Agentic Testing Platform and Practice for Every AI-First Enterprise

Invest in integrated fabric, not the tool:

Start with the layered architecture. Tools are incidental and can change, but the fabric can't.

Calibrate every agent against the NAU-A:

If we can't score and calibrate the agent on V1/V2/V3, we can't define the strategy, plan, and budget, compare validation costs, or predict TCO.

Move performance testing to the left:

ConPass-style soak/endurance tests uncover what point-in- time load tests never can. For agentic systems and workflows, soak tests must be part of the release quality gate.

Run adversarial testing as the most important test phase:

Advest, Garak, and red-team probes belong in the pipeline and not in the quarterly audit windows.

Hybrid stance is important:

A model-agnostic and platform-specific combined strategy is important. Anything else will create either lock-in or compliance debt.

Ensure workflow intelligence early:

Test for the agentic workflow and agent interactions, not just the agent. Failures emerge from A2A interactions.

Make the quality fabric the trust contract:

This needs to be positioned for regulators, boards, and customers to build early trust in agent systems, and this is not just for engineering teams.

Conclusion

Agentic AI has already changed how enterprises operate. It hasn't yet changed how most enterprises assure quality, and that gap is where risk quietly accumulates. Closing this gap starts with establishing a common foundation that keeps cost, performance, and risk manageable as AI adoption grows. It also requires embedding performance and resilience testing throughout the development lifecycle and evaluating end-to-end workflows, rather than assessing individual agents in isolation.

None of this holds unless the fabric is positioned as a trust contract, one regulators, boards, and customers can rely on, not just engineering teams. That is the shift every AI-first enterprise needs to make: treating quality as what keeps autonomous systems safe without constant human oversight.

About the Authors

Anjali Chhabra

Global Head - Practice & Strategic Pursuits,
Agentic Development & Modernization Services,
Tech Mahindra

Anjali Chhabra is a technology leader with 25+ years in IT engineering services, specializing in SDLC optimization, DevSecOps, application modernization, and quality engineering. She has led large enterprise initiatives across BFSI, telecom, RCG, and technology, delivering cloud and on-premises modernization, automated pipelines, and communities of practice.

Pat Sullivan

Expert Advisor, Third Eye Advisory

Pat Sullivan is a veteran technology analyst with more than 25 years of experience advising technology and service providers on market strategy and execution. Recognized by the IIAR as the “Most Influential Analyst of the Year” in 2018, Pat is known for his deep industry insight and pragmatic guidance. Before becoming an analyst, Pat had spent two decades in IT services and consulting, holding leadership roles across several global firms. His perspective blends hands on operational experience with a forward looking view of digital transformation, making him a trusted voice for organizations navigating evolving technology landscapes.

References

1. Gartner. (2025, June 25). Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027. Gartner. 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

2. Singla, A., Sukharevsky, A., Hall, B., Yee, L., Chui, M., & Balakrishnan, T. (2025, November 5). The state of AI in 2025: Agents, innovation, and transformation. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

About Tech Mahindra

Tech Mahindra (NSE: TECHM, BSE: 532755) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 146,000+ professionals across 90 countries, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies.

www.techmahindra.com
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra

Copyright © Tech Mahindra Ltd 2026. All Rights Reserved.

Disclaimer: Brand names, logos, taglines, service marks, tradenames and trademarks used herein remain the property of their respective owners. Any unauthorized use or distribution of this content is strictly prohibited. The information in this document is provided on "as is" basis and Tech Mahindra Ltd. makes no representations or warranties, express or implied, as to the accuracy, completeness or reliability of the information provided in this document. This document is for general informational purposes only and is not intended to be a substitute for detailed research or professional advice and does not constitute an offer, solicitation, or recommendation to buy or sell any product, service or solution. Tech Mahindra Ltd. shall not be responsible for any loss whatsoever sustained by any person or entity by reason of access to, use of or reliance on, this material. Information in this document is subject to change without notice.