Infosys topaz

VIEW POINT

ASSURANCE OF AI AGENTS – BUILDING TRUST AND RELIABILITY

Abstract

Agentic Systems are a transformative technology promising significant automation and efficiency gains across various industries, making quality assurance paramount for their reliable and ethical operation. The agentic AI market size is expected to grow at a rate of 43.61% CAGR from USD 6.96 billion to USD 42.56 billion over the period from 2025 to 2030 (Mordor Intelligence Analysis March 2025). Enterprises are replacing rule-based bots with autonomous agents that manage unstructured, exception-heavy workstreams. Executive surveys show 61% of CEO are integrating Agents into core operations, a level that surpassed the earlier adoption of RPA wave. Diverse sectors, including retail, healthcare, and finance, are at the forefront of exploring and implementing Agentic AI solutions to revolutionize their operations and customer experiences.

This paper underscores the critical importance of integrating quality assurance early in the agentic system development lifecycle. It delves into essential aspects of assurance, including functional correctness, ethical considerations, security protocols, and performance efficiency. Furthermore, it outlines diverse methodologies and approaches to effectively achieve these quality standards.

Infosys®
Navigate your next

1. Insights

- The technological advancements in artificial intelligence and Large Language Models are paving way to the next big wave of agentic systems
- This whitepaper describes the validations needed in various stages of agent development to ensure the quality
- It talks about the various aspects of agentic systems which should be considered for holistic QA
- It tries to propose possible approaches and means to the quality assurance of agentic systems

2. Introduction

Agentic Systems are intelligent entities capable of perceiving their environment, making autonomous decisions, and taking actions to achieve specific goals without constant human oversight. Their autonomy is enabled by:

- **Perception modules** for sensing the environment
- **Reasoning and decision-making units** for processing information
- **Action execution capabilities** to interact with the surroundings
- **Memory or knowledge bases** for learning and adaptation

These systems operate through a continuous plan-think-act cycle: planning objectives, processing information to make decisions, and acting on the environment. Learning from outcomes allows them to improve over time.

Ensuring the quality of autonomous Agentic Systems is a critical imperative as their use expands. A robust quality assurance framework is essential to guarantee reliability, safety, and ethical operation. Our approach involves rigorous testing methodologies tailored to their unique characteristics. This strategy includes:

- **Functional validation**
- **Ethical evaluations**
- **Security assessments**
- **Performance benchmarking**

We use specialized tools and techniques to maintain oversight throughout the system's lifecycle, fostering trust in their dependable deployment.

3. Pillars of Quality Assurance for Agentic Systems

There are various aspects of quality which need to be evaluated in various stages of an AI agent. There should be a systematic way of doing this to address both the operational and ethical issues an AI agent can pose in its lifetime.

- **Functional Correctness**: This pillar focuses on verifying that the agentic system accurately and consistently achieves its intended goals and objectives. It involves rigorous testing to ensure that the agent's planning, decision-making, and actions lead to the desired outcomes under various operational conditions.
- **Ethical Alignment**: Ensuring the agent operates within ethical boundaries and aligns with human values is crucial. This involves evaluating the agent's behavior for fairness, bias, transparency, and accountability, aiming to prevent unintended negative consequences or discriminatory actions.
- **Security and Resilience**: This pillar addresses the need to protect the agentic system from malicious attacks, unauthorized access, and system failures. It includes measures to ensure data integrity, confidentiality, and the agent's ability to recover and continue functioning reliably even in challenging or adversarial situations.
- **Performance and Efficiency**: Evaluating how well the agent utilizes resources (e.g., time, computation, energy) to achieve its goals is essential for practical deployment. This involves assessing the speed, scalability, and resource consumption of the agent processes to ensure optimal operation.
- **Explainability and Transparency**: Understanding why an agent makes specific decisions or takes certain actions is vital for building trust and enabling effective human oversight. This pillar focuses on developing methods to provide insights into the agent's reasoning process, making its behavior more transparent and interpretable.
- **Robustness and Adaptability**: Agentic systems should be able to operate reliably across a range of environments and adapt to changes or unexpected situations. This involves testing the agent's ability to handle novel inputs, recover errors, and maintain performance even when faced with unfamiliar circumstances.

4. Assurance Throughout the Agent Life Cycle

AI Agent Life Cycle

AI agents have a life cycle which spans across its conceptualization to design and deployment

1. Conceptualization and Planning

- Define clear goals
- Break down into subtasks and process flows to achieve the goal
- Plan the resources, processes, and required effort as per the defined goal.

2. Design

- Select suitable agentic frameworks (LangGraph, AutoGen, Google ADK etc.)
- Select LLM as per the complexity of the underlying tasks. For simple tasks simple models will suffice whereas for complex tasks multi-model approaches will be needed.
- Develop the agentic system combining these AI models and tools together with proper decision-making logic and continuous learning.

3. Training and Implementation

- Collect data that the AI agent will use for learning. Then clean and preprocess it to ensure the data is clean, formatted correctly, and free of inconsistencies.
- Use the prepared dataset to train the model by providing the data into the model and further adjust parameters to optimize performance.
- Fine-tune the model by adjusting hyperparameters and retraining to improve performance.

4. Validation

- Integration and system testing
- Performance evaluations and benchmarking
- Stress testing and load testing
- Behavioral validation with edge cases and Adversarial testing

5. Deployment

- Deploying the AI agents into the operational environment
- Continuously monitor the performance of AI agents and record it for future fine tuning of performance
- The AI agents will be deployed in real time environments after complete evaluation and signing off. The performance of AI agents needs to be continuously monitored at this stage. Proper fail-over strategies need to be defined to handle unforeseen agentic failure situations

6. Retirement

- Gradually reduce the AI agent's responsibilities and tasks over time, allowing other systems or newer AI agents to take over.
- Document and archives for future references
- Knowledge transfers to new systems

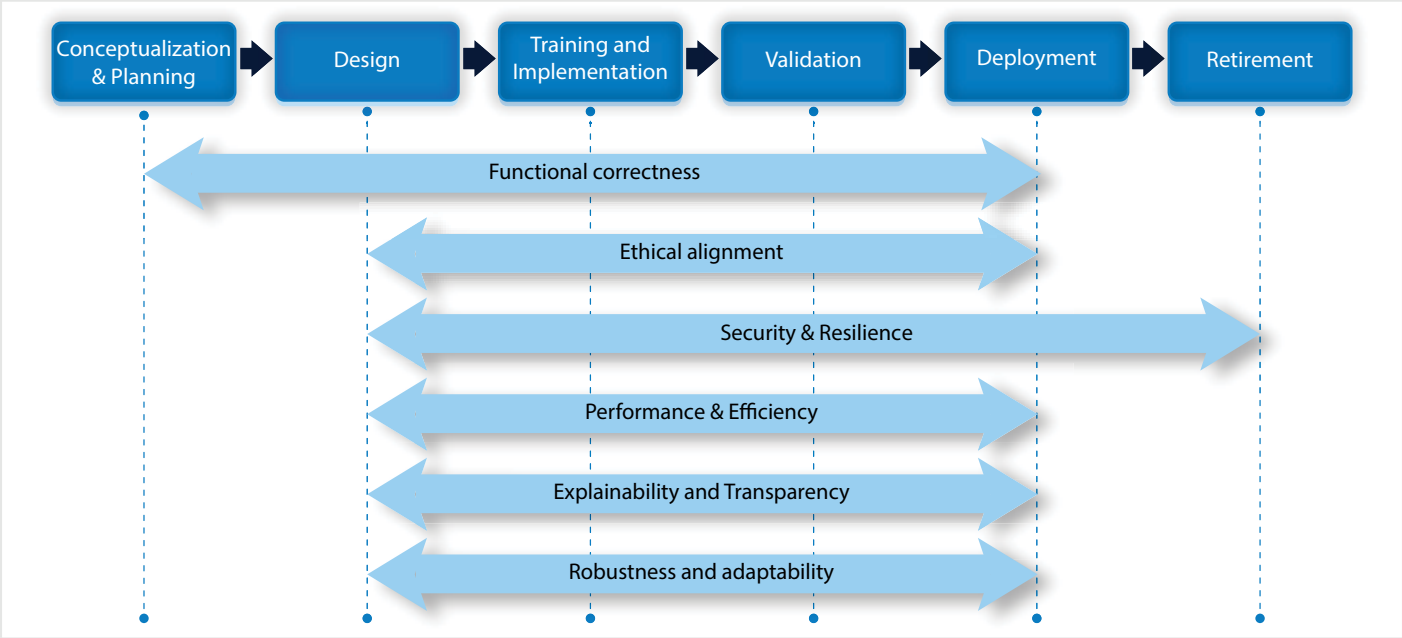

Figure 1. AI life cycle stages and the quality pillars which need to be covered in each of the stages

5. Quality Assurance in AI Agent Life Cycle

Quality assurance should begin from early in the life cycle and span across the life cycle to ensure the agents perform their tasks accurately, efficiently, and ethically.

#	Stage	Testing Type	Approach	Tools
1	Conceptualization & planning	Requirement validation	Manual verification through stakeholder interviews	Possibility of partial automation using NLP to check coverage
		Feasibility analysis	SWOT analysis to assess technical, operational, and financial feasibility along with risks	Nil
		Conceptual model validation	Review high-level AI system design and assumptions	Nil
2	Design	Architecture validation	Architecture reviews	Lucidchart, Enterprise Architect
		Component interface testing	API validations	Postman, Rest Assured, Ready API
		Model benchmarking	Comprehensive model/Agent bench marking	MLPerf, OpenAI Evals, Hugging Face evaluate etc.
		Explainability & interpretability planning	Ensure design includes explainability mechanisms	Framework like SHAP, LIME, etc. Tools like Microsoft InterpretML
		Compliance design testing	Check that design adheres to legal and regulatory standards	GDPR tools, HIPAA compliance tools
3	Training & Implementation	Model training validation	Ensure the model is learning correctly and converging as expected	TensorBoard, MLflow, Weights & Biases
		Data quality testing	Verify training data integrity, completeness, and correctness	Great Expectations, Pandera, Deequ
		Bias and fairness testing	Assess model outputs for bias and ensure fairness across groups	Fairlearn, AI Fairness 360, What-If Tool
		Performance benchmarking	Compare model performance against baseline and benchmarks	scikit-learn metrics, Tensor-Flow Model Analysis
		Robustness testing	Test model stability under noisy or adversarial inputs	TextAttack, Adversarial Robustness Toolbox
		Integration testing	Ensure agent APIs integrates correctly with downstream systems	Postman, Rest Assured, Ready API

#	Stage	Testing Type	Approach	Tools
4	Validation	E2E functionality validation	Automated functional testing, Human-in-the-Loop validation	Playwright, Selenium
		Requirement coverage analysis	Requirement traceability & defect analytics	AI based analytic tools in Infosys Quality Engineering Platform
		Real-world scenario evaluation	Simulate real-world edge cases	Feasible to use Generative AI to simulate real world scenario
		Comprehensive benchmark testing	Evaluate model accuracy and precision on various datasets	Scikit-learn, TensorFlow Model Analysis, MLflow
		Compliance and audit testing	Validate adherence to regulatory and Responsible AI standards	Partially automated with tools like IBM AI Fairness 360, Microsoft Fairlearn etc.
		Security testing	Threat Modeling, Red Teaming, Adversarial Testing, Penetration Testing	OWASP ZAP, Burp Suite, Nessus
		Resilience testing	Chaos engineering, fault injection	Chaos Mesh, Gremlin
5	Deployment	Performance testing	Evaluate system responsiveness and stability under load	JMeter, Locust, LoadRunner
		Deployment validation	Verify successful deployment and configuration	CI/CD pipelines (Jenkins, GitHub Actions), Ansible
		Rollback testing	Test rollback procedures in case of deployment failure	Can be explored further with generative AI for automation
		Environment compatibility testing	Ensure AI agent works across different environments	Can be explored further with agentic framework
		Continuous monitoring	Continuous monitoring of operational phase	Prometheus, Grafana, ELK Stack
		User Acceptance Testing (UAT)	Validate deployment meets user expectations	Mostly by surveys, feedback forms, manual testing
6	Retirement	Data archival testing	Verify that all relevant data is archived securely and accessibly	Database export tools, backup systems, AWS Glacier, Azure Archive Storage
		Audits	Various audits to check decommissioning, knowledge transfer, legal closure, system cleanup	Partially can be assisted with LLMs for audits
		Post-retirement monitoring	Monitor for any residual effects or issues after retirement	Prometheus, Grafana

4. Organization Level Strategies for Assurance of AI Agents

Developing and deploying AI Agents for a business is a significant undertaking, and it requires a distinct set of organizational QA strategies compared to traditional software or human agent QA. The unique challenges of AI (like data dependency, algorithmic bias, ethical considerations, and continuous learning) necessitate a more holistic and integrated approach to quality assurance.

Here are the key organizational QA strategies required to build AI Agents for their business:

- **Governance & Policy**: Organizations should have established policies and procedures for AI agent evaluations to boost the culture of accountability and standard compliance. There are various standards to help define a proper governance model. (1) ISO/IEC 42001:2023 is the international standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations.
- **Establish a Dedicated AI QA Framework & Teams**: Create a specialized organizational structure and methodology for AI QA, including developing or upskilling teams with expertise in data science, machine learning, and AI-specific testing techniques.
- Implement Robust Data Quality Management (DQM): Prioritize and institutionalize comprehensive DQM processes across the organization, covering data collection, cleaning, labeling, and governance to ensure the foundational input for AI agents is accurate, consistent, and unbiased.
- **Integrate Ethical AI & Bias Mitigation Strategies**: Embed organizational policies and technical strategies to proactively identify, measure, and mitigate algorithmic bias at every stage of the AI lifecycle, ensuring fairness, transparency, and accountability in agent behavior.
- **Adopt a Shift-Left & Continuous Testing Paradigm**: Implement a culture and processes that integrate QA activities from the earliest stages of AI development (e.g., problem definition, data selection) and leverage DevOps pipelines for automated, iterative testing.
- **Prioritize Performance, Robustness & Security Testing**: Go beyond functional testing to systematically evaluate AI agents for their performance (latency, throughput), robustness (handling adversarial inputs, edge cases), and security vulnerabilities, ensuring resilience in real-world scenarios.
- **Leverage Human-in-the-Loop (HITL) & Expert Validation**: Design organizational workflows that incorporate human oversight and expert review of AI agent decisions, particularly in critical or ambiguous situations, to provide continuous feedback and improve model accuracy and reliability.
- **Ensure Comprehensive Production Monitoring & Observability**: Establish robust organizational systems for real-time monitoring of AI agent performance in production, tracking key metrics like accuracy, drift, and user feedback to detect and address issues promptly.
- **Establish Rigorous Version Control & Traceability**: Implement organizational standards and tools for comprehensive version control of AI models, datasets, code, and configurations, ensuring reproducibility, auditability, and clear lineage of development.
- **Embrace Explainable AI (XAI) Principles**: Where applicable and critical for business or regulatory reasons, invest in and integrate XAI techniques to help interpret and understand AI agent decisions, building trust and aiding in debugging and validation.
- **Ensure Regulatory Compliance & Audit Readiness**: Develop and maintain organizational processes and documentation to ensure AI agents comply with relevant industry standards and regulations, data privacy laws like GDPR, and ethical guidelines, preparing for potential audits.

By implementing these organizational QA strategies, businesses can not only ensure the technical quality of their AI agents, but also build trust, mitigate risks, and ultimately drive greater business value from their AI investments.

5. Conclusion

In conclusion, assuring AI agents demands a holistic and proactive approach that includes technology, ethical oversight, and organizational alignment. By integrating robust testing strategies covering all QA pillars, organizations can build AI systems that are not only high-performing but also trustworthy and resilient. As AI continues to evolve, a well-structured assurance framework will be essential to ensure these agents operate safely, transparently, and in alignment with human values and regulatory expectations. Commercial and open-source tools are available which can be used effectively by coping with human expertise to cover some level of validations. This presents a compelling opportunity for innovation. With the advent of generative AI, there is now the potential to build intelligent automation tools that can fill these gaps—tools capable of generating test scenarios, simulating edge cases, and even reasoning about ethical implications in ways that were previously infeasible. By harnessing generative AI, we can move toward a future where agent validation is not only more comprehensive but also more adaptive and scalable.

References:

- Agentic AI Market Size & Share Analysis - Industry Research Report - Growth Trends
- "Understanding The 7 Stages of The AI Agent Lifecycle"- Author Shubham Sahu
- "Top Metrics for Evaluating Ethical AI Frameworks" - Dustin W. Stout
- "AI Agent & Agentic AI Survey Statistics 2025 | SS&C Blue Prism" - Alexis Veenendaal
- "National Technology Day 2025: Agentic AI – India's Next Tech Frontier? | Entrepreneur" - Shivani Tiwari
- "https://arxiv.org/pdf/2503.12687" - Naveen Krishnan
- "AI Lifecycle from a Data-Driven Perspective: A Systematic Review" - Wang, Di, Ruiyang Chen, Chuanni Li, and Shanshan Gu. 2025
- "Governing AI Agents"- Noam Kolt

About the Authors

Saji V.S. - Senior Principal Technology Architect, ICETS has 25 years of IT industry experience with over 18 years of experience in Products and Platform development and deployment. He is part of Infosys Center for Emerging Technology Solutions and heading the Autonomous Quality Engineering platform

Jitty Joseph - Senior Technical Manager, ICETS is a Senior Technical Manager in Infosys Center for Emerging Technology Solutions with 19 years of experience in the IT industry. She has worked extensively on quality engineering platforms and IPs designed to ensure the quality and user experience of digital applications.

For more information, contact askus@infosys.com

Infosys®
Navigate your next