INFUSE

A Risk-Centred Framework for Artificial Intelligence Deployment:

The critical role of testing, validation, and governance.



Abstract

The rapid adoption of artificial intelligence (AI), particularly generative and agentic systems, has introduced significant opportunities alongside substantial risks.

While organisations are accelerating deployment to gain competitive advantage, evidence suggests that many deployments are premature, insufficiently governed, and inadequately tested.

This paper argues that AI systems should not be treated as deterministic software artefacts, but as probabilistic, non-deterministic entities requiring continuous validation.

By synthesising insights from recent AI testing practices and emerging deployment risks, this white paper proposes a comprehensive framework for AI testing that integrates behavioural validation, adversarial testing, governance, and lifecycle monitoring.

The findings highlight that robust testing is not merely a quality assurance function, but a practical mechanism for trust, compliance, and safe scalability. It also positions AI assurance as a practical delivery discipline, rather than a purely theoretical governance concern.

Our specialist experience in Higher Education highlights the need to provide guardrails for content, ring-fencing to protect against data leakage, and also content usage monitoring – these should be seen as high-risk categories for AI testing.



Matthew Wheeler,
MD of Partnerships, Infuse

1. Introduction

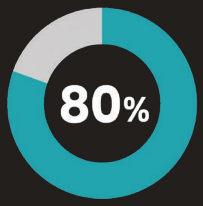
Artificial intelligence adoption has entered a phase characterised by urgency and competitive pressure.

Organisations are increasingly deploying AI systems—particularly autonomous or agentic AI—into real-world workflows without sufficient validation. Evidence indicates that a **significant proportion of organisations have already experienced unintended AI actions**, reflecting systemic gaps in testing and governance.

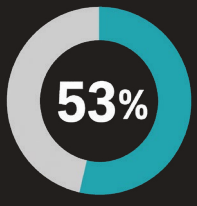
A critical misconception underpinning this acceleration is the assumption that AI systems behave like traditional software systems. In reality, AI systems operate as **"plausibility engines" rather than deterministic "truth engines"**, generating outputs that may appear coherent and confident while being factually incorrect.

This distinction fundamentally challenges conventional software testing methodologies and necessitates new approaches.

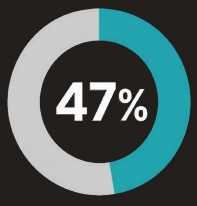
Unintended AI Actions



80% of organisations report AI agents taking unintended actions



53% of organisations have seen AI exceed its intended permissions



47% of organisations experienced AI-related security incidents

Sources: Sailpoint (2025); Cloud Security Alliance (2026). See page 13 for full references.

1.1 The Infuse AI Assurance Delivery Model

Infuse's approach to AI assurance builds on our established in-house processes for project governance and test service delivery: iPMM (Infuse Project Management Methodology) and iSDM (Infuse Service Delivery Methodology).

The methodologies provide a practical delivery framework that can be mapped onto a client's operating model and adapted to local centre of excellence standards, while maintaining the delivery discipline expected from independent frameworks like TMMi. This gives our clients confidence that AI assurance is being delivered through a mature and repeatable quality framework.

For AI-enabled systems, this model is extended to address the specific assurance challenges created by non-determinism, hallucination, behavioural drift, adversarial manipulation, data exposure and agentic autonomy. iSDM provides the delivery lifecycle; iPMM provides the governance, reporting, risk management, change control and closure discipline that supports each stage.

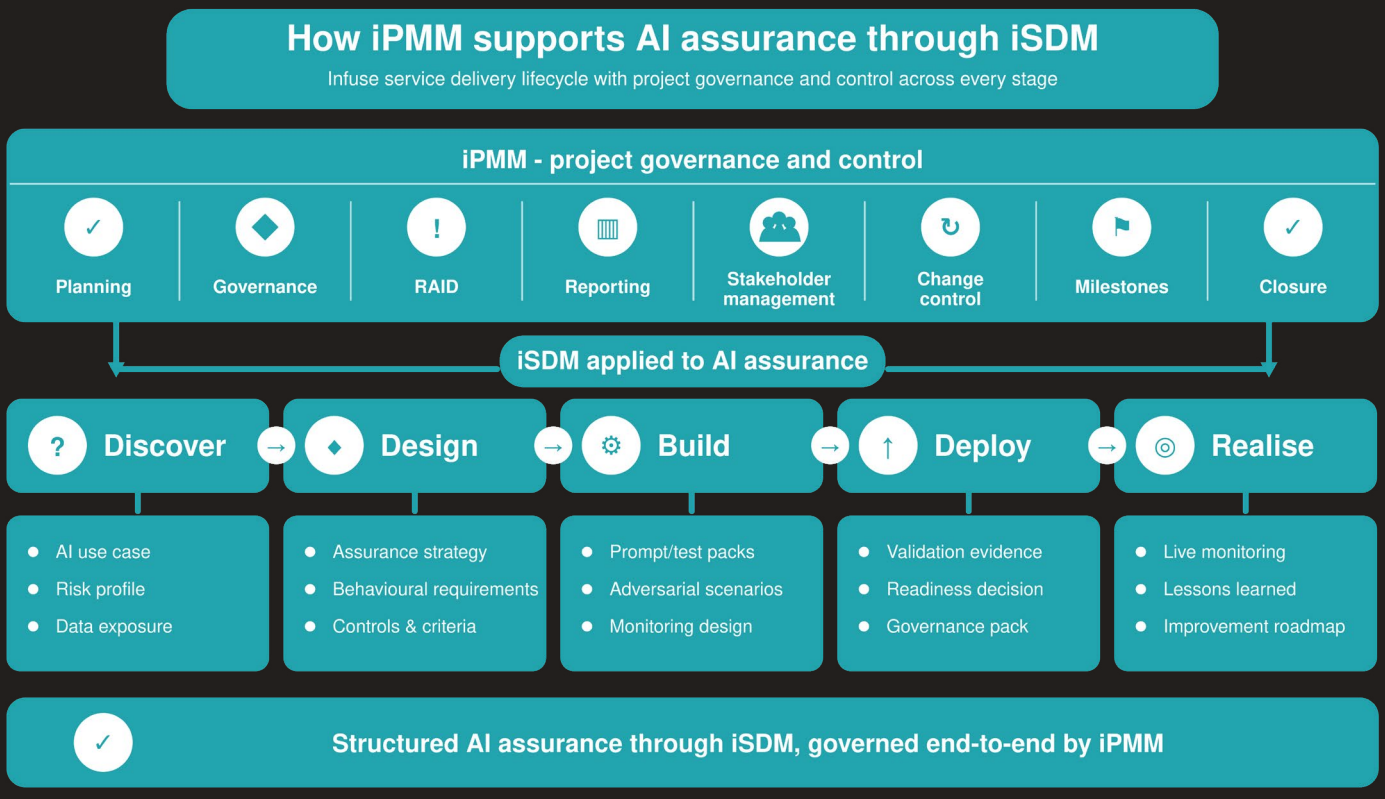


Figure 1: Applying iPMM and iSDM to AI assurance

This framing is important because AI assurance should not be treated as a one-off testing activity. It should move through discovery, design, build, deployment and realisation, with clear evidence, review points and improvement actions at each stage.

2. The Emerging Risk Landscape in AI Deployment

2.1 Non-Determinism and Uncertainty

Unlike traditional systems, AI models produce non-deterministic outputs, where identical inputs may yield different results. This unpredictability undermines conventional pass/fail testing models and introduces ambiguity in validation processes.

2.2 Hallucinations and Hidden Errors

AI systems can generate **plausible yet incorrect information**, often with high confidence. These errors are increasingly difficult to detect and may propagate unnoticed into decision-making workflows. This creates particular risk in sensitive domains such as education, healthcare, legal, and financial services.

2.3 Over-Trust and Cognitive Bias

Users tend to over-rely on AI outputs, especially when presented in fluent and authoritative language. This leads to reduced verification behaviour and increased likelihood of accepting incorrect outputs.

2.4 Security and Data Risks

AI deployments introduce novel vulnerabilities, including:

- Prompt injection and adversarial manipulation
- Cross-session data leakage and exposure of personally identifiable information (PII)
- Misconfigured role-based access controls (RBAC)

These risks extend beyond traditional cybersecurity frameworks, requiring AI-specific security testing.

2.5 Complexity of Multi-Model Architectures

Modern AI platforms often rely on **multiple underlying models**, dynamically routing requests between them. This hidden architectural behaviour can produce inconsistent outputs and obscure accountability.

3. Limitations of Traditional Testing Approaches

Traditional software testing practices were designed for deterministic systems where a given input is expected to produce a consistent and predictable output.

While these approaches remain a critical foundation for assuring application quality, they are not sufficient when applied to AI-enabled systems. Large Language Models (LLMs) and other AI technologies operate probabilistically, meaning that the same prompt may generate different, yet still valid, responses depending on context, model state, and underlying statistical processes.

This fundamentally challenges traditional quality assurance methods that rely on predefined test cases and exact output comparisons.

Furthermore, conventional testing techniques focus primarily on functional correctness and system reliability, offering limited visibility into AI-specific risks such as hallucinations, biased outputs, data leakage, toxicity, or susceptibility to adversarial user inputs. As a result, organisations that apply only traditional testing methods may overlook significant issues that could impact trust, compliance, safety, and user experience.

The limitations become even more pronounced once AI systems are deployed into live environments. Traditional testing is typically performed at fixed points in the development lifecycle and assumes system behaviour remains stable following release.

AI systems, however, can evolve over time through model updates, changing data sources, shifts in user behaviour, and emerging threat vectors. Static validation approaches therefore provide only a snapshot of performance and fail to detect degradation, drift, or newly introduced risks.

In addition, conventional security testing frameworks were not designed to address AI-specific attack surfaces such as prompt injection, jailbreak attempts, indirect prompt manipulation, model extraction, or misuse of generated content.

Consequently, organisations that rely solely on legacy testing practices may develop a false sense of confidence in their AI deployments. Effective assurance requires augmenting traditional quality engineering with specialised AI testing disciplines, continuous monitoring, adversarial evaluation, governance controls, and ongoing risk assessment to ensure AI systems remain safe, secure, reliable, and aligned with organisational objectives throughout their operational lifecycle.

4. A Comprehensive AI Testing Framework

4.1 Continuous Behavioural Validation

Testing must extend beyond pre-deployment into continuous monitoring and validation of system outputs. This should include repeatable regression packs that can be re-run when prompts, models, retrieval sources, connectors or policies change, alongside:

- Real-time output evaluation
- Detection of hallucinations and inaccuracies
- Longitudinal tracking of system behaviour
- Usage monitoring

While testing a conversational AI platform for a university client, we found significant value in reviewing the queries users were submitting. It helped identify misuse patterns, out-of-scope requests and opportunities to improve the overall user experience. As a safeguard, a designated user-group was established to review these inputs, ensuring the wider staff were not exposed to inappropriate or sensitive information.

4.2 Probabilistic and Quality-Based Evaluation

AI validation should shift from exact output matching to **quality-based scoring models**, incorporating:

- Risk-based acceptance criteria aligned to the business impact and autonomy of the AI use case
- Output grading frameworks
- Prompt-response benchmarking

This enables measurable quality assessment despite non-determinism, without relying on arbitrary accuracy percentages that may be inappropriate for high-risk use cases.

4.3 Adversarial and Security Testing

Testing must simulate real-world attack scenarios by:

- Performing prompt injection and manipulation tests
- Stress-testing guardrails through repeated attempts
- Evaluating resilience against harmful or unsafe outputs
- Testing whether the AI can be manipulated into misusing connected tools, data sources or permissions

This approach reflects the necessity to “think like an attacker” during validation.

When testing a conversational AI platform used by a university client, we saw first-hand how easily guardrails can be probed via prompt injection using text, image and document-based payloads. The tests went beyond just text prompts but also used multi-modal injection vectors.

4.4 Data Protection, Equality and Compliance Testing

AI systems must be rigorously tested for:

- PII exposure and data leakage
- Cross-session contamination
- Compliance with data protection regulations (e.g., GDPR)
- Bias and fairness testing, particularly where AI informs admissions, attainment analytics, student support triage or Widening Participation activity
- Evidencing compliance with the Public Sector Equality Duty alongside GDPR, given the high-risk classification of education and vocational training use cases under the EU AI Act

This reduces legal and reputational risk.

Remember, data leakage is not always the result of malicious activity, it can occur unintentionally during normal AI usage. For example, a member of staff may upload a large Excel spreadsheet containing sensitive student information to ask the AI to reformat, analyse, or summarise the data.

If appropriate controls are not in place, this could result in sensitive information being processed by a model or service in a way that creates a risk of exposure to unauthorised individuals or unintended data retention. This highlights the importance of user awareness, data handling policies, and technical safeguards alongside data protection/security testing.

During a project testing an enterprise AI assistant, we wanted to see whether the model could be steered into discussing out-of-scope or sensitive subject matter, separate from pure prompt injection. For example, whether the assistant consistently stays within its intended domain and adheres to policy, regardless of how repeatedly the user phrases the request.

Bias and fairness testing should be treated as a core component of AI assurance rather than a compliance exercise. AI systems do not intentionally discriminate, but they can reproduce, amplify or conceal biases present in training data, retrieval sources, prompt design, system instructions or institutional processes. In Higher Education, these risks become particularly significant where AI-generated outputs may influence decisions affecting students, applicants or staff.

Effective fairness testing should therefore examine not only whether protected groups receive statistically different outcomes, but also whether the AI system behaves consistently when presented with equivalent scenarios involving different demographic characteristics. This may involve testing variations of prompts, personas, case studies and user journeys to identify unintended differences in recommendations, prioritisation, language, tone or access to information.

Testing should also recognise that fairness cannot be reduced to a single metric. An AI system may satisfy one measure of fairness while creating inequitable outcomes elsewhere. As a result, assurance activities should combine quantitative assessment with human review and domain expertise, particularly when evaluating use cases involving admissions, student support, academic progression, wellbeing, conduct management or widening participation initiatives.

Where AI is used to support decision-making, organisations should be able to evidence how fairness risks have been identified, assessed and mitigated. This includes maintaining clear records of testing activities, identified biases, corrective actions and residual risks. Human oversight remains essential, particularly where AI outputs could materially affect an individual's opportunities, experience or outcomes.

In our experience, fairness testing often reveals wider organisational issues rather than purely technical defects. Models may simply reflect inconsistencies, historical biases or gaps that already exist within source content and business processes. Consequently, bias and fairness assurance should be viewed as an organisational learning activity, helping institutions improve both their AI systems and the governance frameworks surrounding them.

For universities, this approach supports compliance with equality obligations while also strengthening trust amongst students, staff and external stakeholders. More importantly, it helps ensure that AI-enabled services enhance inclusion and opportunity rather than unintentionally reinforcing existing inequalities.

4.5 Multi-Model and System-Level Testing

Testing should target not only individual models but also **platform-level behaviour,** including:

- Dynamic model routing
- Output consistency across models
- Hidden architectural decision-making

4.6 Performance and Scalability Validation

Given the scale of AI deployments, testing must include:

- Simulated high-volume usage scenarios
- Stress testing of token and session management
- Identification of backend limitations

Our performance testing exposure on an enterprise AI assistant demonstrated that AI may interpret repeated traffic differently, so the system may not actually be under the load volume we always assume. Conventional scenario generation and load profiling approaches do not translate directly, and this needs its own methodology.

Traditional performance testing assumes that a given input reliably produces a comparable unit of backend work, allowing volume and load profiles to be modelled with confidence. AI systems break this assumption: the same request can be interpreted, expanded, or routed differently depending on context, prompt history, retrieval results, or which underlying model handles it.

As a result, repeated or seemingly identical traffic may not generate consistent backend load, and the system may not actually be under the volume of load that test scenarios intend. Token consumption, response length and processing time can all vary significantly between runs, even where the test inputs themselves are unchanged.

This means AI performance testing needs to account for behavioural variability as a first factor, ensuring that load profiles reflect realistic variation in model behaviour rather than assuming the deterministic, repeatable demand patterns that conventional performance testing relies upon.

4.7 Human-in-the-Loop Oversight

Critical decisions should incorporate **human validation layers**, particularly in high-risk contexts. This ensures accountability and mitigates the impact of incorrect outputs.

5. Governance and Organisational Implications

5.1 AI as a Governed Digital Entity

AI agents should be treated as controlled digital identities, subject to governance, access control, monitoring and revocation frameworks. Once AI can access systems, retrieve documents, trigger workflows or recommend actions, it becomes an operational actor and should be governed accordingly.

5.2 Testing as a Driver of Requirements Maturity

AI testing frequently uncovers undefined or incomplete requirements, highlighting its role as a design and discovery mechanism rather than purely validation.

5.3 Collaboration and Stakeholder Alignment

Effective AI testing requires continuous communication between technical teams, business stakeholders, and governance functions to ensure relevance and adoption of findings.

5.4 Academic Integrity and Assessment Assurance

The rapid adoption of generative AI presents both opportunities and challenges for academic integrity. While AI can enhance learning, improve accessibility and support student productivity, it also creates new risks for assessment validity, authorship verification and academic misconduct. For universities, the challenge is not simply detecting AI use, but ensuring that assessment processes continue to measure the knowledge, skills and competencies they are intended to evaluate.

AI assurance should therefore include dedicated testing of academic integrity controls and assessment-related use cases. This involves evaluating how AI systems interact with institutional assessment policies, learning environments and student workflows. Universities need confidence that AI-enabled services support legitimate learning outcomes without creating pathways that undermine assessment authenticity or disadvantage particular groups of students.

Testing activities should consider a range of scenarios, including:

- The generation of assessment answers, coursework drafts and reflective submissions
- The ability of students to use AI tools to circumvent intended learning objectives
- The effectiveness of disclosure, citation and attribution mechanisms where AI-assisted work is permitted
- The consistency of AI-generated feedback, marking support or assessment recommendations.
- Potential inequalities created by differential access to AI tools and technologies
- The risk of students being incorrectly accused of misconduct based on unreliable AI-detection approaches

Particular attention should be paid to assessment design. In many cases, assurance activities reveal that vulnerabilities arise not from the AI system itself but from assessment formats that were developed before widespread AI adoption.

Testing can help institutions identify assessment methods that are highly susceptible to AI-assisted completion and evaluate alternative approaches such as authentic assessments, oral examinations, practical demonstrations, portfolio-based evidence and iterative coursework development.

Where AI is used to support marking, feedback generation or academic decision-making, universities should validate that outputs remain aligned with institutional regulations, assessment criteria and academic standards.

Testing should examine consistency, transparency and explainability, ensuring that academic staff retain accountability for final decisions and that AI outputs cannot inadvertently introduce bias, factual inaccuracies or inconsistent judgements.

Our experience within higher education suggests that assessment assurance should focus on the broader educational outcome rather than attempting to eliminate AI use entirely.

The objective is to ensure that students continue to demonstrate genuine learning and achievement in a manner that is fair, transparent and defensible.

As AI capabilities continue to evolve, academic integrity assurance should become an ongoing activity, combining policy, governance, assessment redesign, user education and technical testing to maintain confidence in institutional awards and qualifications.

6. How to Scale Safely

A key paradox emerges from the analysis: organisations that succeed in AI adoption are not those that move fastest, but those that create enough delivery discipline to scale safely. This approach enables:

- Increased trust in AI outputs
- Reduced risk exposure
- Sustainable scalability and continuous improvement

Premature deployment, by contrast, results in the propagation of errors, erosion of trust, and potential regulatory breaches.

Risk-Based AI Assurance and Deployment Triage

A common misconception is that all AI systems require the same level of testing, governance and oversight. In practice, applying identical assurance processes to every AI use case can unnecessarily delay innovation, increase costs and create governance fatigue. The objective should not be to slow down all AI adoption, but to apply assurance activities in proportion to the potential impact of failure. This aligns testing effort with organisational risk appetite and enables lower-risk use cases to be deployed more rapidly while concentrating assurance resources on areas that could result in significant harm.

We recommend adopting a risk-triage model that assesses AI systems against three primary dimensions:

Impact of Incorrect Output
What would happen if the AI generated an incorrect, misleading or incomplete response?
- Minimal impact: inconvenience or minor productivity loss.
- Moderate impact: operational disruption requiring correction.
- Significant impact: material business, financial, regulatory or reputational consequences.
- Critical impact: potential harm to individuals, legal exposure or major institutional impact.

Level of Autonomy
How much authority has been delegated to the AI?
- Advisory only.
- Recommendation support.
- Automated execution with human approval.
- Fully autonomous action.

Sensitivity of Data
What information does the system process, access or generate?
- Public information.
- Internal operational data.
- Confidential institutional data.
- Personal, sensitive or regulated data.

AI Assurance Triage Matrix

Risk Level	Example Use Cases	Assurance Requirements
Tier 1: Low Risk	Meeting summaries, document drafting, internal productivity assistants, knowledge search	Standard security review, basic prompt testing, usage monitoring
Tier 2: Moderate Risk	Student enquiry assistants, staff support chatbots, policy guidance tools	Behavioural testing, security testing, governance review, performance validation
Tier 3: High Risk	Student support triage, wellbeing recommendations, admissions support, attainment analytics	Comprehensive AI testing, bias and fairness assessment, data protection review, human oversight requirements
Tier 4: Critical Risk	Automated decisions affecting enrolment, progression, disciplinary action, funding, safeguarding or legal outcomes	Executive governance approval, independent assurance, continuous monitoring, mandatory human decision-making authority

Risk-Based Delivery Principles

Under this model, assurance effort increases as business impact and autonomy increase. A generative AI assistant helping staff draft meeting notes should not face the same governance burden as an AI system influencing admissions decisions or student support interventions.

For example, a university knowledge chatbot that retrieves publicly available policy information may be suitable for accelerated deployment following focused testing of accuracy, security and guardrails. Conversely, an AI service used to prioritise students for wellbeing support or identify individuals considered at risk of non-continuation should be treated as a high-risk use case requiring fairness testing, extensive validation and human oversight.

This approach enables institutions to:

- Accelerate low-risk innovation.
- Focus assurance investment where it delivers the greatest value.
- Avoid unnecessary governance overhead.
- Demonstrate proportional control to auditors and regulators.
- Build organisational confidence through evidence-based deployment decisions.

The Infuse AI Assurance Principle

Move fast where failure is tolerable.
Move carefully where people, reputation, compliance or institutional outcomes are affected.

The goal of AI assurance should not be to create barriers to adoption, but to provide a structured mechanism for deploying AI at the speed appropriate to its risk. By combining risk-triage with proportionate testing and governance, organisations can accelerate innovation while maintaining trust, accountability and control.

7. Conclusion

AI deployment represents a fundamental shift in how software systems are developed, tested, and governed.

The inherent non-determinism, opacity, and scalability of AI systems require a redefinition of testing from a static quality assurance function to a dynamic, continuous risk management discipline.

This paper demonstrates that:

- AI testing must integrate behavioural, adversarial, and probabilistic methodologies
- Continuous validation is essential to maintain trust and accuracy
- Governance frameworks must evolve to treat AI as an active operational entity

Ultimately, the success of AI adoption will not be determined by speed, but by the ability to test, validate and govern AI systems effectively before and during deployment, moving from promising pilots to controlled, evidenced and repeatable adoption.

Without such practices, organisations risk deploying not intelligence, but uncertainty at scale.

For clients, the practical challenge is to turn experimentation into controlled, repeatable and governable operational AI capability.

References:

SailPoint (2025) SailPoint research highlights rapid AI agent adoption, driving urgent need for evolved security. Available at: https://www.sailpoint.com/press-releases/sailpoint-ai-agent-adoption-report (Accessed: 6 October 2026).

Nuth, T. (2026) The State of Cloud and AI Security in 2026. Cloud Security Alliance, 13 March. Available at: https://cloudsecurityalliance.org/blog/2026/03/13/the-state-of-cloud-and-ai-security-in-2026 (Accessed: 6 October 2026).

About Infuse

Since 2010, our unrivalled sector expertise and end-to-end testing capability have been helping UK universities go faster and further for less, elevating quality, agility and the student experience. With skills rarely found in a single supplier, we bring something no sales pitch deck can replicate — a trusted, proven reputation for delivering results for the UK higher education sector.

info@infuse.it

INFUSE

we are the experts in higher education software testing