Grant Thornton

AI Validation in practice

Real-world use cases for governance, risk and compliance

Executive summary

AI systems are rapidly moving from experimental tools to core components of business operations - particularly in risk, and compliance functions. As adoption accelerates, institutions face need to evolve beyond conceptual governance frameworks and demonstrate how AI systems can be validated in practice.

In our previous paper, *Validating Artificial Intelligence Systems: A Modern Capability for Model Risk Management*, we outlined a structured validation framework tailored to the characteristics of Generative and Agentic AI systems. This paper builds on that foundation by demonstrating how such a framework can be applied in real-world settings, using practical insights drawn from business applications of AI solutions. It highlights:

- How AI validation activities can be translated into specific, testable controls;
- Where AI systems introduce distinct risk types beyond traditional models;
- Why scenario-driven and behavioural testing are essential; and
- How validation outputs support governance, oversight, and regulatory readiness.

The key takeaway is clear - validating AI systems requires dynamic, system-level assurance that is grounded in expert-led insights, measurable outcomes, and verifiable conclusions.

From framework to practical application

As outlined in our previous publication, the current generation of AI systems extends beyond predictive modelling. These systems generate narrative outputs, retrieve and synthesise information dynamically, and, in many cases, execute multi-step workflows supported by reasoning and tool interaction. In doing so, they introduce a level of behavioural variability that challenges conventional validation approaches.

AI validation framework

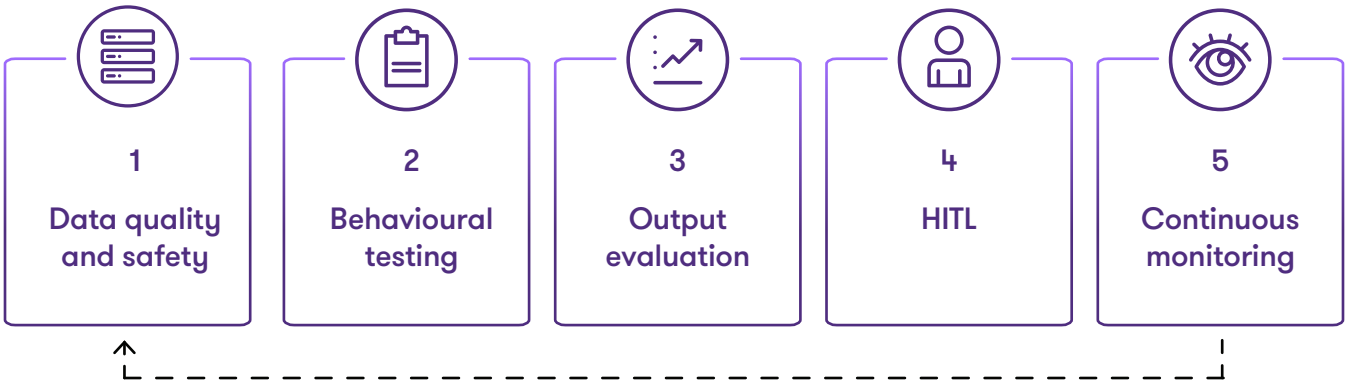


Figure 1. End to end validation framework for Generative and Agentic AI (Source)[1]

Our framework addresses this shift by extending traditional model risk practices to incorporate behavioural and operational considerations. While the core principles of model risk management remain unchanged - including the need for robustness, transparency, and independent challenge - their application must evolve to adequately reflect the characteristics of AI-driven systems.

In practice, this means that validation is no longer a single point-in-time assessment, but a structured process that evaluates the system dynamically across its full lifecycle. We examine two case studies on the application of our AI Validation Framework to AI Agents being adopted by financial institutions, specifically:

- A regulatory compliance assessment agent; and
- A documentation and reporting agent.

Whilst the AI agents present a unique set of challenges for validation, we illustrate how our validation framework can be effectively applied to address AI specific risks, in a way that is familiar to model risk practitioners and senior stakeholders.

Each case study will present an illustrative, non-exhaustive set of tests that can be performed across the components of our validation framework, with further discussion on interpretation of testing results.

[1] *Validating Artificial Intelligence Systems: A Modern Capability for Model Risk Management*

Case study 1: Regulatory compliance assessment agent

As the burden of demonstrating compliance with an increasing volume of regulatory requirements grows, financial institutions are adopting AI agents to accelerate regulatory interpretation and alignment.

Problem statement

Financial institutions are required to interpret and apply large volumes of regulatory requirements across their operations, often relying on manual processes and expert judgment. These approaches can be resource-intensive, and may lead to inconsistencies, particularly where regulatory requirements are principles-based or subject to interpretation.

AI solution

A Reg Compliance Assessment Agent is trained to support the interpretation and analysis of specific regulatory content. The agent ingests information from approved sources, including regulatory publications and internal documentation, and processes this information to identify key obligations and assess alignment of existing policies and controls. It can then identify potential compliance gaps and remediation actions.

Intended outcomes

An effective AI reg compliance assessment agent can deliver;

- **Efficient regulatory Rreviews** – Reduces time and resource requirements associated with identifying and achieving regulatory compliance.
- **Improved policy and control alignment** – Helps identify gaps or inconsistencies between regulatory requirementsand existing policies and frameworks.
- **Scalability** – Handles ever-growing volumes of regulatory content and internal documentation without a proportional increase in cost and effort.

Associated challenges

Regulatory texts are often complex and context-dependent, requiring nuanced interpretation that may not always be fully captured by an AI system. There is also a risk of over-reliance on AI-generated outputs, particularly where results are presented with a high degree of clarity and structure. Outputs can vary depending on how inputs are phrased and context is provided. This makes it difficult to rely only on fixed test cases or standard performance measures. Instead, validation should include a mix of scenario and input tests, expert review, and structured checks to assess whether outputs are accurate and appropriate.

Ensuring explainability and traceability is critical, particularly in demonstrating how conclusions have been reached and how regulatory requirements have been interpreted. As regulatory expectations evolve and the system is exposed to new scenarios, continuous monitoring is required to ensure that performance remains appropriate and that emerging risks are identified and addressed.

1.1. Validation approach

1.1.1. Data quality & input review

Validation begins with checking that inputs are accurate, complete, and appropriate for the task. This includes regulatory publications, internal policies, and supporting documents. It is important to ensure that outputs are based on approved sources and that there is a clear link between inputs and results.

Illustrative tests:

- **Input completeness check** – Confirm that all required regulatory documents and supporting materials are included before the assessment is run.
- **Input format consistency check** – Ensure inputs follow a consistent structure, such as standard tem-plates or predefined formats.
- **Missing or incomplete input check** – Identify gaps in inputs, such as absent policy sections or unclear requirements.

Working example:

Input completeness check

Purpose:
Confirm that all required regulatory documents, policies, and supporting materials are present before the assessment is executed. Prevent incomplete inputs from leading to partial, misleading, or incorrect compliance conclusions.

Inputs required:
• Document inventory / expected document list
• Uploaded document set
• Mapping of documents to regulatory requirements.

Metric:
Input Completeness Rate = (Number of required inputs confirmed as present, current, accessible and usable / Total required inputs)

Verification method:
Compare the available inputs against the predefined inventory and assess whether each required input is present, current, accessible and usable. Identify missing, incomplete, outdated or incorrectly formatted inputs.

1.1.2. Behavioural testing

Behavioural testing examines how the tool interprets unclear or conflicting regulatory requirements, and whether it escalates or flags uncertainty appropriately. Testing is carried out using different inputs and repeated runs to examine consistency/stability.

Illustrative tests:

- **Consistency test** – Run the same regulatory scenario multiple times with variations in wording to assess consistency of outputs.
- **Conflict handling test** – Assess how the tool responds where regulatory requirements conflict or are unclear.
- **Incomplete input test** – Test how the tool responds to missing or ambiguous inputs, including whether gaps or uncertainty are highlighted.

Working example:

Consistency test

Purpose:
Confirm that the Assessment Agent produces consistent outputs when assessing the same regulatory scenario multiple times. Prevent variations in wording or prompt structure from resulting in different interpretations, conclusions, or identified compliance gaps.

Inputs required:
• Regulatory scenario
• Prompt variations representing the same regulatory requirement
• Internal policies and supporting documentation

Metric:
Consistency Rate = (Number of materially consistent outputs / Total number of outputs generated)

Verification method:
Run the same regulatory assessment multiple times using different prompt wording while keeping the underlying regulatory requirement unchanged. Compare outputs to assess consistency of regulatory requirement, findings, recommendations and conclusions.

An output is materially consistent where prompt variation does not change the identified obligations, compliance gaps, recommendations, risk assessment or overall conclusion.

1.1.3. Output evaluation

Output evaluation focuses on whether responses are accurate, relevant, and suitable for use. This includes checking that outputs are supported by source material. Outputs are also reviewed against the original task to ensure they stay on topic and cover all key points.

Illustrative tests:

- **Source traceability check** – Verify that material conclusions can be linked back to specific regulatory sources or policy documents.
- **Completeness check** – Confirm that all required sections (e.g. requirements, gaps, recommendations) are present.
- **Usability check** – Assess whether outputs can be used directly in reporting or require further clarification.

Working example:

Source traceability check

Purpose:
Assess whether material interpretations and conclusions are supported by specific approved sources and accurately represent those sources. This prevents unsupported conclusions, ensures transparency in analysis, and allows outputs to be reviewed, challenged, and audited effectively.

Inputs required:
• Source regulatory texts (e.g. guidance, standards, rules)
• Internal policies and control frameworks
• Agent output (requirements, gaps, conclusions)

Metric:
Traceability rate = (Material claims supported by appropriate source evidence / Total material claims tested)

Verification method:
Trace each material interpretation and conclusions to the relevant source provision and confirm that the source has been represented accurately and in context.

Where applicable, assess whether references are sufficiently specific (e.g. paragraph level rather than general document level).

1.1.4. Human-in-the-Loop (HITL)

Human review is a key part of validation, where outputs are examined by regulatory experts at critical points where judgement is required. This ensures that outputs are appropriate and that final decisions remain with the user.

Illustrative tests:

- **Escalation trigger review** – Identify cases where outputs require additional review due to unclear or disputed interpretations.
- **Expert judgement check** – Apply human judgement to validate interpretations of regulatory requirements.
- **Final review control** – Ensure outputs are reviewed and approved before being used in decision-making.

Working example:

Escalation trigger review

Purpose:
Identify scenarios where outputs require additional human review before use. This ensures that complex, unclear, or high-impact regulatory interpretations are not relied upon without appropriate oversight, reducing the risk of incorrect conclusions or misinterpretation.

Inputs required:
• Outputs containing regulatory interpretations
• Cases with incomplete or unclear information
• Scenarios involving conflicting or high-risk regulatory requirements

Metric:
Escalation accuracy rate = (Number of correctly escalated outputs / Total outputs requiring escalation)

Verification method:
Review outputs across different scenarios to assess whether appropriate escalation triggers are applied.

Assessment should focus on whether escalation is applied consistently and appropriately, rather than excessively or insufficiently.

1.1.5. Continuous monitoring

Validation continues after initial testing through ongoing monitoring of outputs over time. In practice, this includes checking output consistency, reflecting updated regulatory guidance and internal policies and identifying issues as they arise.

Illustrative tests:

- **Change impact test** – Review outputs following updates to regulatory guidance or input data.
- **Consistency monitoring test** – Track outputs over time to identify changes in consistency or quality.
- **Recurring Issue Tracking** – Identify repeated issues (e.g. missing requirements or unclear outputs) and supports updates to inputs or guidance.

Working example:

Regulatory change impact test
Purpose: Assess how the Assessment Agent responds to updates in regulatory requirements and ensure that outputs reflect the most current rules and guidance. This prevents reliance on outdated information and ensures that assessments remain relevant, accurate, and aligned with current regulatory expectations. **Inputs required:** • New or revised regulatory guidance • Existing policies and control frameworks • Outputs generated prior to and following regulatory updates **Metric:** Regulatory alignment rate = (Number of outputs correctly reflecting updated requirements / Total outputs impacted by regulatory change) **Verification method:** Compare outputs generated before and after a regulatory update to assess whether affected requirements, gaps, recommendations and conclusions are accurately updated.

Case study 2: Documentation and reporting agent

As the burden of demonstrating compliance with an increasing volume of regulatory requirements grows, financial institutions are adopting AI agents to accelerate regulatory interpretation and alignment.

Problem statement

Firms generate a significant volume of documentation across reporting, assurance, and client-facing operations. These involve multiple inputs, structured analyses, and presenting conclusions in a clear and consistent manner.

Documentation is typically time-intensive and reliant on individual judgement, which can result in variability in quality, consistency and professional standards.

AI solution

A Documentation and Reporting AI Agent is designed to produce structured narrative outputs, such as reports, summaries, and commentary, based on source documents, data extracts, evidence framework, and user prompts.

The agent organises inputs into a clear narrative and shapes the output in line with a predefined format. It operates in a supporting capacity, providing a draft which can then be reviewed and refined by users prior to finalisation.

Intended outcomes

Robustly designed Documentation and Reporting Agents support:

- **Streamlined Drafting** – Initial drafting time is reduced, allowing greater focus on interpretation, review, and refinement.
- Structured Articulation of Reasoning – Logical presentation of an analysis, ensuring that key assumptions, rationale, and conclusions are clearly and consistently expressed.
- **Standardisation of Deliverables** – A uniform approach to how documentation is structured and presented, reducing reliance on individual writing styles and improving alignment with established formats.
- **Faster Iteration and Refinement** – More efficient revision of outputs, allowing content to be quickly adjusted in response to feedback or adapted for different audiences.

Associated challenges

The use of AI for documentation and reporting introduces risks primarily related to the quality and reliability of narrative outputs. Issues can arise where outputs are not fully aligned to source material, or where reasoning lacks coherence, particularly in cases of incomplete or ambiguous inputs. Inconsistency across tasks may occur despite similar objectives, reflecting differences in structure or level of detail.

In more complex cases, risks may extend to how outputs are structured and developed, where gaps or inconsistencies emerge in the flow of analysis. There is also a risk of over-reliance, where outputs are accepted without sufficient review. These risks are addressed through validation, which focuses on assessing whether outputs are accurate, relevant, appropriate, and suitable for their intended use.

2.2. Validation approach

Due to the open-ended nature of generative outputs, it is important to establish appropriate controls for the Documentation and Reporting Agent to balance automation with human oversight. These measures ensure that the agent can operate efficiently while maintaining oversight, traceability, and accountability for the final outputs.

2.1. Data quality & input review

This step focuses on whether inputs, including source documents, data extracts, analytical notes, and user prompts, are clear and well-formed to support reliable outputs.

Illustrative tests:

- **Input quality check** – Focus on completeness, clarity, and format by validating input structure against expected requirements.
- **Content safety check** – Screen text for prohibited content, unsafe instructions, or embedded prompts that could influence interpretation or alter the intended output.
- **Sensitive data handling** – Assess the presence of sensitive or restricted information, with such content flagged or redacted prior to use.
- **Prompt formulation control** – Review how instructions are framed to ensure consistent input structures.

Working example:

Input completeness check

Purpose:
Assess whether prompts and instructions are framed clearly, consistently, and in a way that supports reliable outputs from the AI system. This helps reduce variability in responses, prevents ambiguity in user inputs, and ensures that the system receives the information required to perform the intended task accurately.

Inputs required:
• Prompt templates
• System instructions
• User input examples
• Output samples
• Internal policies or assessment criteria

Metric:
Prompt effectiveness rate = Number of tested prompts producing outputs aligned with defined task and output requirements / Total prompts tested

Verification method:
Review prompt templates and sample inputs to assess whether they clearly define the task, context, required inputs and expected output. Test prompts across different users, scenarios and input variations to determine whether outputs remain consistent and fit for purpose, and whether incomplete inputs are appropriately identified.

2.1.1. Behavioural testing

Behavioural testing focuses on how the agent performs under different conditions, rather than evaluating a single output in isolation. Validation at this stage centres on a set of structured tests:

Illustrative tests:

- **Consistency test** – Repeat execution of the same task, including the use of paraphrased or equivalent prompts to detect contractions, semantic drift, or unstable conclusion across repeated runs.
- **Ambiguity and missing data test** – Scenario-based testing using incomplete, ambiguous, or conflicting inputs.
- **Guardrail and refusal behaviour test** – Evaluate whether the agent correctly refuses or limits responses when inputs are outside scope, incomplete, or conflict with predefined constraints.

Working example:

Consistency test

Purpose:
Assess whether the AI agent applied predefined guardrails appropriately when faced with input that are outside scope, incomplete, unsafe, unsupported, or inconsistent with its operating constraints. This helps ensure that the agent does not generate misleading, unauthorised, or inappropriate outputs, and that it responds in a controlled and policy-aligned manner.

Inputs required:
• Guardrail rules
• System instructions
• Internal policies
• Test prompts
• Expected refusal or limitation criteria

Metric:
Guardrail adherence rate = Number of responses correctly refusing, limiting, escalating, or requesting clarification / Total guardrail-triggering scenarios tested

Verification method:
Test the agent using prompts designed to trigger predefined guardrails and assess whether the response is appropriate, consistent, and aligned with internal requirements.

This includes checking whether agent refuses requests that fall outside its authorised scope, confirming that incomplete or ambiguous inputs result in clarification requests rather than unsupported conclusions, assessing whether the agent avoids generating outputs that conflict with policy, control requirements, or system instructions, validating that refusal responses are clear, proportionate, and explain the limitation without providing prohibited or unsupported content. Where applicable, guardrail behaviour should be tested across different scenarios, user phrasings, and escalation attempts to confirm that controls are applied consistently.

2.1.2. Output evaluation

Output evaluation focuses on whether generated content is reliable and suitable for its intended use.

Illustrative tests:

- **Format reliability test** – Validate the outputs against expected structures, such as predefined sections, tables, or report templates.
- **Factual integrity test** – Compare outputs against underlying source material to confirm that claims are supported and traceable.
- **Hallucination review** – Review to identify and quantify unsupported or fabricated content to detect where facts, figures, and conclusions are introduced without evidence.

Working example:

Input completeness check

Purpose:
Assess whether the agent introduces unsupported, inaccurate, or fabricated content in its outputs. This helps ensure that conclusions, facts, figures, references, and recommendations are grounded in the provided input data, approved sources and clearly stated assumptions. The agent should be able to appropriately signal uncertainty where evidence is limited.

Inputs required:
• Internal policies
• Generated outputs
• Approved reference materials
• Evidence or citation requirements

Metric:
Hallucination rate = (Number of unsupported or fabricated claims identified / Total factual claims reviewed)

Verification method:
Review generated outputs against the underlying input data and approved source materials to determine whether each material claim is supported by evidence. This includes reviewing whether assumptions are clearly labelled and do not appear as confirmed facts, and whether there are any unsupported claims or fabricated references.

Completeness check – Evaluate against task requirements to confirm that all required elements are present and no critical information has been omitted.

2.1.3. Human-in-the-Loop (HITL)

Human-in-the-loop review focuses on the incorporation of human judgement as a validation control, particularly where outputs are material, ambiguous, or require interpretation.

Illustrative tests:

- **Materiality-based review trigger** – Apply predefined criteria to route outputs for review based on characteristics such as complexity or ambiguity and verify that escalation occurs consistently.
- **Override and editing pattern analysis** – Compare draft and final versions across outputs to identify recurring user corrections and systematic weaknesses in the agent.
- **Accountability validation test** – Review workflows to confirm that outputs cannot be finalised without defined human review.

Working example:

Materiality-based review trigger

Purpose:
Assess whether the agent applies predefined materiality and risk-based criteria to identify outputs that require additional review, challenge, or escalation. This helps ensure that higher-risk inputs, including those involving complex judgement, ambiguity, significant impact, or limited supporting evidence, are appropriately routed for human review before being relied upon.

Metric:
Escalation accuracy rate = (Outputs correctly routed for review / Total outputs meeting the trigger criteria)

Verification method:
Review a sample of generated outputs against the predefined materiality and review trigger criteria to assess whether escalation occurred appropriately and consistently.

2.1.4. Continuous monitoring

Ongoing monitoring is necessary to keep outputs within acceptable bounds as inputs, usage patterns, and expectations evolve over time.

Illustrative tests:

- **Output drift detection test** – Compare outputs over time for similar prompts to identify changes in structure, interpretation, or reasoning patterns.
- **Consistency trend monitoring** – Track variability in outputs for equivalent prompts across repeated runs and over time to detect instability or degradation.
- **Editing effort trend test** – Compare draft and final outputs over time to track changes in the level of human refinement required.
- **Trigger-based revalidation test** – Monitor changes in prompts, inputs, or models, and initiate targeted re-testing where material changes occur.

Working example:

Input completeness check

Purpose:
Assess whether the agent's outputs remain stable and consistent over time when responding to similar prompts, inputs, or assessment scenarios. This helps identify output drift, including changes in structure, interpretation, reasoning approach, tone, or conclusions that may affect reliability, comparability, or alignment with intended use.

Inputs required:
• Approved baseline prompts and outputs
• Current comparable prompts and outputs
• Consistent input data or source material
• Expected output requirements
• Relevant change and monitoring logs

Metric:
Output stability rate = (Number of outputs remaining consistent with baseline expectations / Total comparable outputs tested)

Verification method:
Compare baseline and current outputs using equivalent prompts and inputs. Assess whether material changes in structure, reasoning, conclusions or recommendations are explainable by approved changes to the model, prompts, source information or system configuration. Unexplained material variation should be flagged for investigation.

Where applicable, testing should be performed across multiple scenarios, use cases, and time periods to confirm that output behaviour remains stable and that drift is identified promptly.

Setting and interpreting validation thresholds

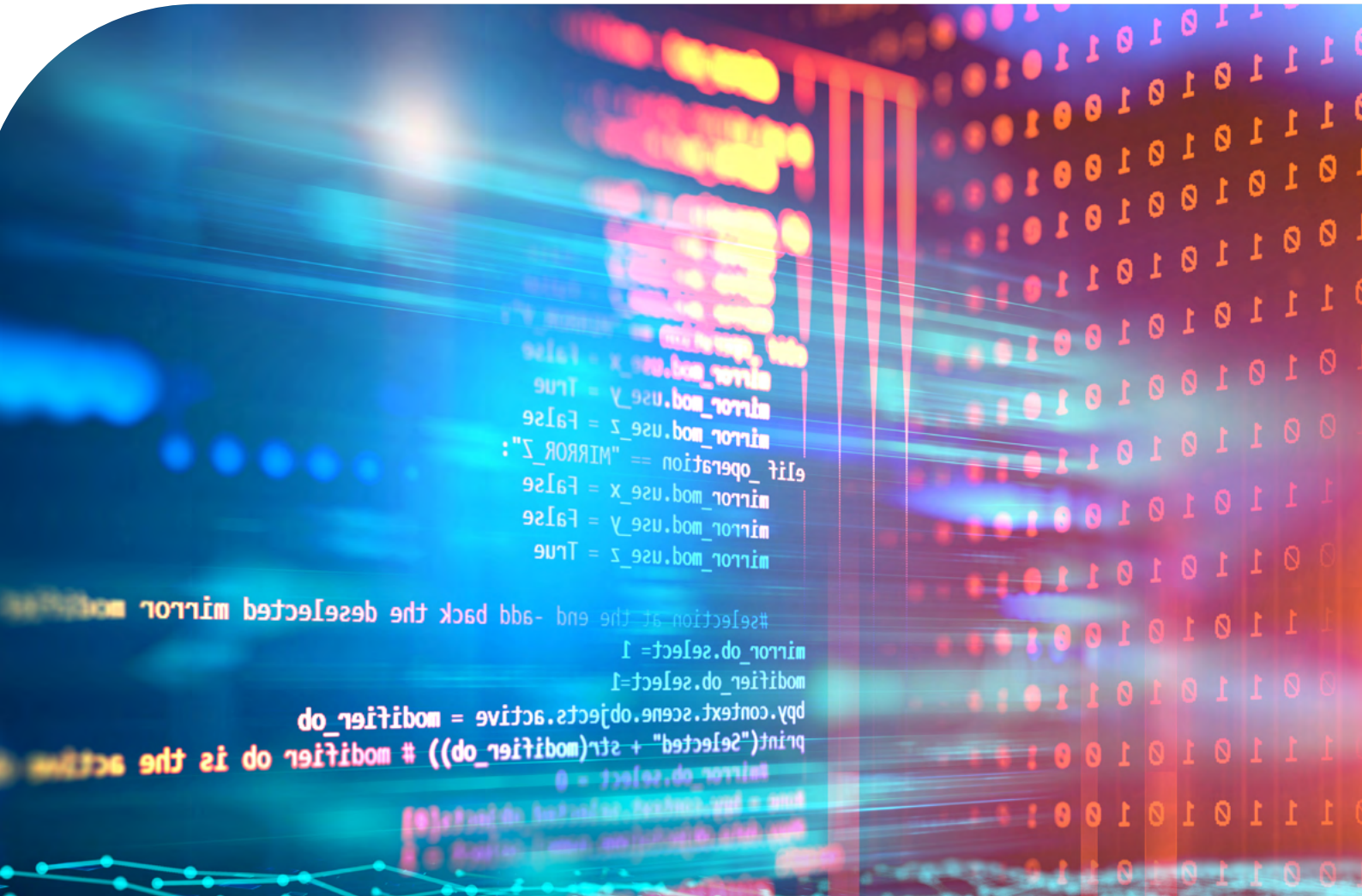
Validation thresholds should be calibrated to the intended use, risk profile and control environment of the AI system. They should not be treated as universal benchmarks applicable to all tests and use cases.

Quantitative thresholds may be appropriate where outcomes are objectively measurable, such as input completeness, source traceability or escalation performance. For tests involving judgement, such as the regulatory change test or prompt formulation control, results should be assessed based on the materiality and potential impact of identified exceptions.

The following general criteria can inform RAG ratings:

- **Green:** No material exceptions are identified and the result supports reliance within the validated scope.
- **Amber:** Limitations are identified but are non-material, appropriately controlled or capable of remediation without preventing controlled use.
- **Red:** A material weakness affects the reliability, traceability, oversight or controlled use of the AI system, or results in risk outside the organisation's tolerance.

Ratings should not be based on the number of exceptions alone. A single material failure may result in a Red rating, while several minor exceptions may remain Amber. Thresholds should be more stringent where AI outputs support regulatory interpretation, customer outcomes, financial decisions or other high-impact activities. Metrics and thresholds should inform, rather than replace, expert validation judgement.



Conclusion

The case studies discussed above demonstrate that effective AI validation must extend beyond accuracy testing. It should assess the suitability of inputs, the system's behaviour across realistic scenarios, the reliability and traceability of outputs, the effectiveness of human oversight, and whether performance remains appropriate over time.

Several practical implications emerge:

- Validation should reflect how the AI system is actually used: Testing should cover realistic variations in prompts, inputs and operating conditions, including ambiguity, incomplete information and conflicting instructions.
- Accuracy alone is insufficient: An output may appear plausible while remaining incomplete, unsupported or inconsistent. Validation should therefore assess source grounding, completeness, stability and suitability for the intended purpose.
- Materiality matters more than the number of exceptions: A single unsupported regulatory interpretation, fabricated material claim or missed escalation may be more significant than several minor formatting or drafting issues.
- Human oversight must be demonstrably effective: Human review should not be treated as a general safeguard unless the circumstances requiring review, the responsible reviewers and the evidence of challenge and approval are clearly defined.
- Metrics support, but do not replace, expert judgement: Quantitative measures can provide useful evidence, but validation conclusions should also consider the severity of exceptions, their potential impact, the effectiveness of mitigating controls and the residual risk remaining.
- Validation does not end at deployment: Changes to models, prompts, source information, system configuration or intended use can alter system behaviour. Ongoing monitoring and trigger-based revalidation are therefore necessary to maintain confidence over time.

Taken together, these activities provide an evidence base for determining whether an AI system is sufficiently reliable, controlled and governed for its intended use. Validation findings should support practical decisions on approval, remediation, usage restrictions, human oversight and monitoring, rather than simply recording whether individual tests have passed.

However, independent validation of this depth cannot reasonably be applied to every AI system across an organisation. The scope, depth and frequency of validation should be proportionate to each system's classification and risk tier. Our next article will explore how AI system classification and risk tiering can be used to determine an appropriate level of validation and assurance for AI systems.