International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Adapting the Five Pillars of Model Risk Management for Generative AI: The GEN-5 Validation Framework

Dr. Nabanita Sinha
Associate Director, Deloitte, India

ABSTRACT: The emergence of Generative Artificial Intelligence (AI) systems has expanded the boundaries of traditional model development and validation, introducing new dimensions of model risk. Existing Model Risk Management (MRM) standards such as SR 11-7 and SS1/23 remain foundational; however, their application must evolve to address the dynamic and context-dependent behaviour of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) architectures, and multi-agent environments. These systems pose novel and high-impact risks due to their generative nature, contextual variability, and ability to self-orchestrate actions. Ensuring that such models produce consistent, auditable, and risk-mitigated outcomes has become critical, yet validators face growing challenges in assessing conceptual soundness, monitoring reasoning reliability, and evaluating control effectiveness within existing MRM frameworks.

This paper proposes GEN-5, a five-pillar validation and assurance framework that adapts established Model Risk Management (MRM) principles to the unique behaviours and risks of Generative AI, RAG pipelines, and multi-component AI systems. GEN-5 provides a standardized template and actionable methodology for assessing conceptual soundness, performance accuracy, outcome reliability, control effectiveness, and continuous monitoring across AI-driven environments. It integrates both qualitative and quantitative evaluation techniques—including hallucination detection, prompt robustness, retrieval fidelity, semantic consistency, and reasoning stability—while emphasizing the essential role of governance, safety guardrails, and control assurance. By extending traditional MRM rigor to modern AI architectures, GEN-5 offers practitioners a policy-aligned, technically grounded approach for identifying, evaluating, and mitigating the novel risks introduced by Generative and enterprise-scale AI use cases.

KEYWORDS: Model Risk Management (MRM), GenAI, Large Language Models (LLMs), RAG, Model Validation

1. INTRODUCTION

Model Risk Management (MRM) has long served as the foundation for ensuring transparency, accountability, and reliability in the use of analytical and predictive models across regulated industries. Supervisory standards such as the U.S. Federal Reserve's SR 11-7 [1], the Bank of England's SS1/23 [2], and the Monetary Authority of Singapore's (MAS) [3] collectively define expectations for model governance, validation independence, and ongoing performance monitoring. These policies emphasize three core objectives: establishing sound conceptual design, maintaining robust performance and monitoring practices, and implementing strong governance and control mechanisms throughout the model lifecycle.

However, the rapid proliferation of Generative and Agentic AI systems—driven by Large Language Models, Retrieval-Augmented Generation (RAG) architectures, and multi-agent orchestration—has fundamentally changed the nature of what constitutes a "model." These AI systems differ from traditional statistical or machine-learning models in that they are non-deterministic, context-adaptive, and capable of reasoning or acting autonomously. As a result, they introduce new categories of risk not fully covered by existing regulatory guidance. Examples include hallucination, prompt sensitivity, context contamination, output inconsistency, data leakage, and control bypass through autonomous agents or tool calls.

While SR 11-7 and SS1/23 [1,2] remain structurally relevant, their interpretation must evolve to accommodate the generative paradigm. Traditional validation steps—such as parameter back-testing or benchmark accuracy assessment—are insufficient for models whose outputs are generated dynamically and influenced by prompts, retrieved context, or agent collaboration. The MAS 01/23 [3] Fairness, Ethics, Accountability, and Transparency (FEAT) principles reinforce this need by emphasizing traceability, explainability, and human oversight in AI-enabled systems. For Generative AI, this translates to validating prompt integrity, retrieval relevance, response consistency, and control effectiveness in addition to core model accuracy.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Furthermore, as enterprises automate decision processes using agentic systems—where multiple AI agents plan, reason, and execute tasks—the scope of validation expands beyond a single model to the entire reasoning workflow. Validators must ensure that autonomous or semi-autonomous agents operate within defined boundaries, produce consistent and auditable outputs, and comply with institutional control frameworks. The challenge for MRM teams lies in quantifying qualitative risks such as hallucination or reasoning drift while maintaining compliance with existing regulatory expectations.

To bridge this emerging gap, this paper introduces GEN-5, a five-pillar validation and assurance framework that extends the established MRM principles to the domain of Generative AI. GEN-5 provides a standardized, policy-aligned template to guide model validators in assessing conceptual soundness, outcome reliability, governance effectiveness, and continuous monitoring within AI-driven environments. The framework complements existing supervisory policies—SR 11-7, SS1/23, and MAS 01/23—by offering practical validation metrics and control tests tailored to the dynamic, generative behaviour of modern AI systems.

The remainder of this paper is structured as follows. Section 2 reviews the background and related work, including gaps in current AI assurance practices. Section 3 presents the proposed GEN-5 Framework and explains how its five pillars extend traditional MRM to GenAI, RAG, Multi Agent systems. Section 4 concludes the paper by summarizing the key contributions and implications for AI validation practice.

2. BACKGROUND AND RELATED WORK

The validation of AI systems has evolved significantly as models have shifted from deterministic statistical constructs to dynamic, generative, and context-aware systems. While regulatory guidance such as the U.S. Federal Reserve's SR 11-7, the Bank of England's SS1/23, and the Monetary Authority of Singapore's (MAS) FEAT and 01/23 [1-3] guidelines remain the global benchmarks MRM, these policies were formulated primarily for conventional machine learning models with fixed parameters and well-defined data dependencies. They emphasize conceptual soundness, ongoing monitoring, governance, and independent validation—principles that remain essential but are inadequate for the adaptive, reasoning-driven nature of Generative and Agentic AI systems.

To bridge this regulatory gap, several global initiatives have been launched to standardize AI governance. The NIST AI Risk Management Framework (AI RMF, 2023) [4] outlines a structured approach to identify, measure, and mitigate AI-related risks through governance, mapping, measurement, and management functions. Similarly, the OECD AI [5] Principles and the European Commission's Ethics Guidelines [6] for Trustworthy AI articulate core dimensions such as transparency, fairness, accountability, and human oversight. At an organizational level, the recently introduced ISO/IEC 42001 (2023) [7] establishes the first formal management system standard for AI, providing operational controls for lifecycle governance, traceability, and documentation. Collectively, these frameworks define the "what" of responsible AI, but they stop short of prescribing the "how"—especially in relation to the validation and assurance of LLM-based or agentic systems under enterprise MRM governance.

In the academic domain, several researchers have begun to address the unique failure modes of LLMs and RAG pipelines. Recent studies by Huang et al. (2023) and Yu et al. (2024) [8-10] provide systematic taxonomies of hallucination issue and propose benchmark datasets to quantify factual faithfulness and response reliability. Emerging toolkits such as HalluLens [11], RAGAs, and eRAG [12-13] demonstrate that validation must now assess both retrieval precision and generative faithfulness (e.g., factual consistency, attribution correctness). These works collectively highlight the necessity for dual-component validation, where both the retrieval layer and generative layer are independently tested and then evaluated as an integrated system.

Beyond academia, industry practitioners are also shaping the evolving field of AI validation. Institutions such as Wells Fargo [15] and Citi [16] have publicly discussed their approaches to responsible AI assurance, focusing on bias control, explainability, and human oversight within generative applications. Similarly, consulting and risk advisory firms including Deloitte (2024) [14] and PwC (2023) [17] have published white papers emphasizing the operationalization of model validation for GenAI, introducing structured red-teaming, prompt-injection testing, bias mitigation, and model lineage tracking as essential control activities [18,19]. These industry perspectives reinforce the consensus that traditional validation techniques such as back-testing or sensitivity analysis must be expanded to include prompt robustness, context relevance, and safety filter performance.

This synthesis of academic, regulatory, and practitioner perspectives reveals a clear gap: MRM teams lack a standardized, actionable framework to validate Generative and Agentic AI systems within existing supervisory guidelines. This paper addresses that gap

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

through the proposed GEN-5 Framework—a five-pillar structure that aligns regulatory requirements with modern AI assurance practices, integrating governance rigor, quantitative outcome testing, and control validation specific to the generative paradigm.

3. PROPOSED SOLUTION

The GEN-5 Framework extends the traditional expectations of Model Risk Management (MRM) under SR 11-7, SS1/23 [1-2], and the MAS AI Risk Management Guidelines (AIRG 2025) [23] to the domain of Generative. While foundational MRM principles—conceptual soundness, performance validation, governance, and monitoring—remain relevant, the emergence of LLM-based, RAG-enabled, and multi-agent architectures requires additional validation dimensions. GEN-5 Framework provides a structured approach for validating Generative and Muti Agent AI systems. It organizes validation activities into five focused pillars that help practitioners assess conceptual integrity, behavioural reliability, control effectiveness, and lifecycle risks of AI models.

3.1 GEN 5 Models:

Generative AI (GenAI) refers to standalone large language models that generate outputs purely from their internal learned representations, making them suitable for tasks such as information answering, summarization, content drafting, and reasoning. Typical examples include systems like ChatGPT for general queries or Microsoft Copilot for email generation and document summarization. RAG [12] extends this capability by combining an LLM with an external retrieval layer, enabling the model to ground its responses in enterprise or regulatory documents. RAG is used in applications such as internal policy assistants that retrieve HR manuals, and compliance or regulatory Q&A systems.

In contrast, Prescribed Multi-Agent Systems (PMAS) consist of multiple AI components or task-specific LLM modules that perform specialised steps within a fixed, rule-driven workflow. These components pass information sequentially, based on predefined orchestration logic rather than autonomous reasoning or agentic behaviour. PMAS architectures are commonly used in operational processes such as KYC automation—where one module extracts data, another validates entities, and a third computes risk scores—or AML alert workflows where modules summarise transactions, check red flags, and draft investigator notes [14]. PMAS should not be confused with agentic systems, as they do not plan, negotiate, or make decisions independently; instead, they execute deterministic steps defined entirely by the workflow design.

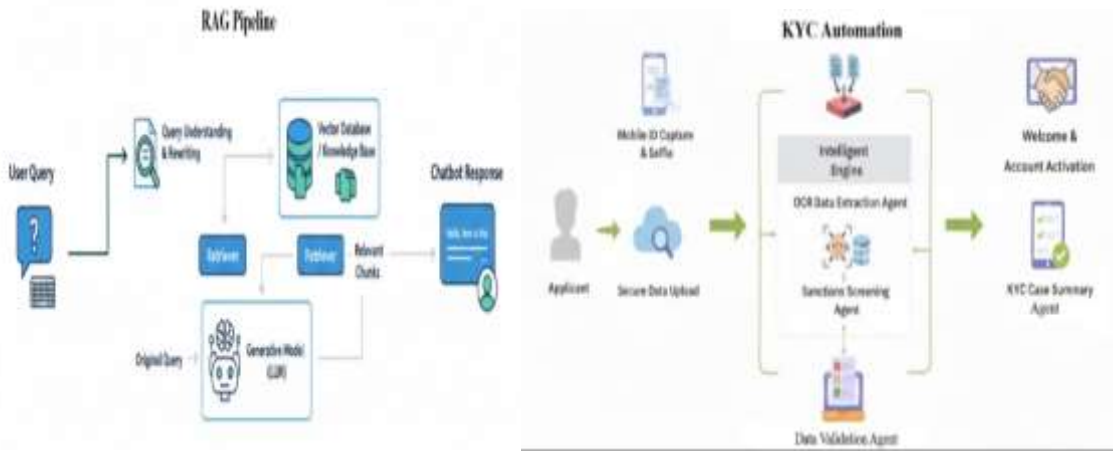

Fig 1. RAG Pipeline & Muti Agent KYC Automation

Figure 1 illustrates a RAG pipeline integrated with an LLM, alongside a KYC automation workflow represented as a prescribed multi-agent system.

3.2 Pillar 1 - AI Model Overview

The first step of validating any GenAI system is to establish a complete and accurate Model Overview. Before testing or assessing behaviour, the validator must confirm that all foundational elements of the model are clearly documented, traceable, and aligned with approved use. This ensures that the validator fully understands what the model is, what it is designed to do, and the environment in which it operates. Pillar 1 must focus broadly on four areas:

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

- ***Model Type and Technical Classification*:** The validator must confirm that the model type is correctly identified, including whether it is a GenAI model, RAG pipeline or PMAS. The technical nature of the model must also be explicitly classified—whether it uses pre-trained foundation models, fine-tuned or LoRA-adapted models, open-source LLMs (e.g., Llama, Mistral), or third-party vendor/API-based models (e.g., OpenAI, Gemini, Claude) [19]. This classification is critical because each category introduces different dependencies, risks, validation requirements, and limitations.
- ***Purpose, Scope, and Intended Use:*** The validator must check that the model documentation explicitly describes: the business purpose and decision context, the intended users and stakeholders, the scope of use, including assumptions and boundaries [24], including the type of usage (e.g., individual productivity tool, organisation-wide productivity application, business decision support, or client-facing service). Validators must also verify the portfolio or population to which the model is applied and ensure the system's operating boundaries, assumptions, and interaction channels are clearly defined.
- ***Ownership, Roles, and Governance Structure:*** The validator must verify that model ownership and accountability are clearly defined. This includes model owner and sponsor, roles for development, testing, deployment, responsibilities for monitoring and issue escalation, approval authorities and governance committees [24]. Clear governance provides accountability for all lifecycle stages.
- ***Dependencies and Architectural Context:*** The validator must ensure that the model documentation describes all key dependencies that influence behaviour or risk [18], including external LLM APIs, embedding models, RAG retrievers and vector databases, agent orchestration or tool-calling frameworks, upstream data sources and downstream decision systems. Understanding dependencies is critical because GenAI and Multi Agent systems often rely on components outside direct model control.

3.3 Pillar 2 - Conceptual & Architectural Soundness

Pillar 2 evaluates whether the design, architecture, and reasoning flow of the AI system are conceptually sound [2], technically coherent, and properly documented. For Generative AI, RAG and Multi-agent systems, conceptual soundness extends beyond mathematical formulation. It must demonstrate why the chosen AI approach is appropriate, how each component works, how inputs flow through the architecture, and how the system maintains safety, relevance, and reliability. Validation under this pillar must cover the following dimensions.

3.3.1 ***Model Architecture and Technical Components:*** Pillar 2 assesses whether the AI system's design, architecture, reasoning flow, and component interactions are technically sound, logically structured, and justified. Before evaluating architecture, the validator must clearly understand what type of AI system it is. Different architectures require different validation expectations. The architectural comparison presented in Table 1 (GenAI vs. RAG vs. Multi-Agent Systems) highlights that Generative AI systems can differ vastly in their internal components, data flows, processing capabilities [21]. Because of these architectural differences, the risks, failure modes, and required control structures vary significantly between models. Therefore, the validator must not treat all GenAI systems uniformly; instead, validation must be tailored to the architecture type.

3.3.2 ***Validation Requirements for GenAI:*** In the following sections, we describe what the validator must examine for each architecture type, and why each step is essential to ensuring conceptual and architectural soundness. Validation Requirements It is essential to understand whether the model is API-based (e.g., GPT-4, Claude, Gemini) or open-source (Llama, Mistral). This matters because: API-based models introduce data privacy and external dependency risk, Open-source models raise model governance and security patching responsibilities, and Fine-tuned models introduce data quality and bias amplification risks [15]. The validator must assess whether the chosen LLM aligns with the organisation's risk posture and whether the rationale for its selection is justified.

3.3.3 ***Validation Requirements for RAG Systems:*** RAG systems introduce retrieval pipelines [9], and therefore the validator must examine both retrieval and generation.

- Embedding Model Appropriateness: Embedding models convert text into vectors; poor embeddings lead to poor retrieval. Validators must understand the embedding model used (i.e. E5, BGE, Ada, OpenAI) and assess whether it is suitable for the domain, capable of capturing regulatory/legal text, multilingual where required. If the embedding model is inappropriate, the entire RAG pipeline collapses.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

- Chunking Strategy: Chunking affects semantic integrity. Overly large chunks cause irrelevant responses; overly small chunks cause incomplete context. Validators must ensure chunking is: consistent, representative of document structure, and appropriate for policies, legal texts, manuals. Chunking errors are the single most common RAG failure mode.
- Vector Database: The validator must review the type of vector database used (e.g., FAISS, Milvus, Pinecone, Weaviate) and the underlying index structure (HNSW, IVF, Flat), as these determine retrieval speed, accuracy,
- Metadata and Access Control: Metadata dictates which documents can be retrieved. Validators evaluate whether metadata includes jurisdiction, department, date, sensitivity level. Poor metadata leads to wrong-source retrieval (e.g., retrieving old MAS guidelines instead of the latest version), which creates regulatory risk.
- Retrieval Logic & Temperature: Validators must examine whether retrieval is using top-k searches, hybrid (BM25 + vector), reranking using cross-encoders and check temperature value. Retrieval logic determines quality and relevance, which directly impacts correctness of outputs.

Many RAG pipelines include pre-retrieval components such as input classifiers, document-type detectors, or query-rewriting modules that operate before the retrieval step [21]. Validators must assess the correctness and reliability of these components because they directly influence what content is searched and how the user query is interpreted. Input classifiers determine whether the user's question is in-domain, which document collection should be queried, and whether the query should be routed, bypassed, or escalated. Query decomposition modules may split complex questions into sub-queries to improve retrieval accuracy. Misclassification or incorrect reformulation at this stage can cause the system to retrieve irrelevant documents or omit critical context, resulting in highly confident but incorrect answers. Therefore, validators must independently test classification accuracy, routing logic, domain detection, and the safety of bypass or escalation rules to ensure that upstream preprocessing does not introduce hidden failure modes into the RAG pipeline.

3.3.4 ***Validation Requirements for PMAS:*** Prescribed multi-agent systems consist of multiple LLM modules or components executing in a fixed, predefined sequence. These systems do not perform autonomous reasoning, planning, negotiation, or arbitration. Their behaviour is fully dictated by the workflow design and orchestration logic [20]. Validation must therefore focus on the correctness, robustness, and auditability of the workflow, not on cognitive agent behaviour.

- Role and Task Specialisation Each component (LLM module, extractor, validator, classifier, scorer) must have a clearly defined and documented role. Ambiguity in module responsibilities leads to duplicated processing, inconsistent outputs, or missing steps. Validators must confirm that each module performs only its intended transformation, without hidden side effects or unintended functionality.
- Hand-off Logic and Workflow Orchestration: Because prescribed systems follow deterministic pipelines, hand-off correctness is critical. Validators must evaluate whether each component receives the expected input format, whether outputs from one step are correctly transformed into inputs for the next, whether failure at any step is surfaced rather than silently suppressed, and whether mandatory validations occur at the correct stage. Breakdowns in hand-offs cause silent downstream failures, making workflow integrity a key validation area [20].
- Tool / API Calling Logic: Prescribed systems may call tools or databases, but these calls are fixed and rule-based, not chosen by autonomy. Validation must ensure correct tool selection per rule, correct parameter passing, robust error handling for failed calls, and clear fallback or escalation when upstream tools return incomplete or malformed data. Any error in tool/API sequencing may cause cascading workflow failures [20].

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Table 1: GenAI, RAG and Multi-Agent - Architecture Overview

Aspect	GenAI (LLM-Only)	RAG Pipeline	PMAS
Purpose	Summarization, drafting, reasoning, or Q&A without external grounding	Context-grounded answers using enterprise documents, regulatory texts, and internal policies	Automating multi-step workflows using fixed logic and task-specific LLM components
Core Components	Single LLM (pre-trained, fine-tuned, or API-based)	Embedding model + vector DB + retriever + LLM	Multiple LLM modules/components orchestrated in a fixed sequence (not reasoning agents)
Inputs	User query (plus optional context shared explicitly by the application)	User query + retrieved chunks + metadata	Predefined task inputs + intermediate outputs passed deterministically between components
Knowledge Source	Internal learned representations of the LLM	Indexed enterprise knowledge retrieved at run-time	Rule-based access to tools, databases, and upstream workflow outputs; no autonomous knowledge choice
Pre-Retrieval / Input Classification Logic	NA	It may include domain classifiers, document-type detectors, routing logic, query rewriting or query decomposition modules.	Usually not required; pipeline follows fixed task logic and does not redirect based on classification
Tool / API Calling	NA	Optional (re rankers or search utilities)	Yes — but entirely pre-scripted and fixed; no autonomous tool selection
Inter-Module Communication	N/A	N/A	Sequential hand-offs only; no negotiation, no reasoning-based interaction
Workflow Nature	Single-step or conversational flow	Retrieval → filtering → generation	Fixed, deterministic workflow defined by rules or orchestration logic (not by the LLMs themselves)
Key Strength	High fluency, creativity, general reasoning ability	High factual consistency and transparency through grounding	High determinism, predictable outcomes, straightforward auditability
Key Vulnerability	Hallucination; ungrounded responses	Retrieval errors, stale data, metadata leakage	Error propagation; brittle logic; no self-correction or adaptive reasoning
Output Characteristics	Fluent, generalizable answers	Grounded, referenceable answers with citations	Deterministic outputs driven by workflow design rather than autonomous reasoning
Risk Level	Low	Medium–High	Medium–High

3.3.5 ***Input Data Quality Check:*** LLMs and most foundation models are pre-trained, meaning the validator is not assessing training data quality, but rather the quality of the input data supplied at inference time, which varies by use case. Therefore, input data quality validation must focus on the actual runtime inputs that drive the behaviour of the GenAI system. For Pure GenAI applications, this includes free-text user inputs and structured prompts, which must be checked for completeness, clarity, formatting, and neutrality.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

For RAG systems, the validator must examine all inputs originating from the vector database, including document preprocessing quality, accuracy of text extraction (from PDFs, tables, scanned images), metadata correctness, and chunking integrity, as these directly influence retrieval relevance and grounding. When the system relies on upstream preprocessing steps—such as image processing, OCR, table extraction, or entity recognition—the validator must assess whether these components introduce errors, distort content, or reduce semantic fidelity.

For PMAS, additional categories of inputs must be reviewed, including intermediate tool outputs, API responses, database lookups, search results, planner scratchpads, and agent memory state. Because agentic behaviour depends heavily on these dynamic inputs, poor quality or inconsistent intermediate data can lead to incorrect plans, unsafe tool calls, or faulty multi-step reasoning. In multi-agent architectures, the validator must also ensure quality and consistency of inter-agent messages as corrupted or incomplete state transfers may propagate errors across agents.

3.4 Pillar 3 - Performance Validation & Outcome Reliability:

Pillar 3 evaluates whether the AI system delivers correct, stable, safe, and reproducible outputs across a wide range of operating conditions. It covers both Performance Validation, which focuses on measurable accuracy and quality against reference standards, and Outcome Reliability, which focuses on stability, consistency, robustness, grounding, and behaviour under stress or ambiguous perturbations [18]. Together, these two components ensure that model outputs are not only accurate under ideal conditions but also trustworthy under real-world uncertainty.

As part of Pillar 3, validators must first assess whether the model development documentation clearly specifies the intended accuracy targets, evaluation datasets, test methodologies, thresholds, safety constraints, and failure-handling logic. Table 2 summarises the core performance metrics commonly used to evaluate the accuracy and semantic quality of GenAI models, while Table 3 highlights the reliability tests that are essential for ensuring stability, robustness, and safety in real-world GenAI use cases

Table 2: Performance Validation Metrics

Metric	Definition	Why It Matters	How to Test
Accuracy & Task-Specific Correctness	Measures correctness vs. ground truth (exact match, F1, precision/recall, structured output match).	Baseline measure of factual and functional correctness for the intended task.	Use labelled datasets; evaluate predictions vs. ground truth; compute accuracy/F1/etc.
Semantic Quality Metrics (BLEU, ROUGE, BERT Score, Cosine Similarity)	BLEU = lexical overlap; ROUGE = recall of key content; BERT Score = semantic similarity; cosine similarity = embedding-level closeness.	Evaluates how well generated text preserves meaning, structure, and fidelity compared to reference answers.	Compute BLEU/ROUGE/BERT Score on validation sets; compare embedding similarity using cosine distance.
Grounding & Attribution Quality (RAG)	Measures how well model outputs align with retrieved context, Metrics: RAG Attribution Score, Citation Accuracy, Grounding Score	Ensures factual grounding and prevents unsupported or hallucinated claims in RAG systems.	Compare outputs to retrieved documents; test retrieval using labelled relevance sets; validate source citations.
Human Evaluation (Qualitative Measures)	Human SMEs judge outputs for correctness, completeness, contextual fit, clarity, tone, and compliance.	Captures nuances and failure cases automated metrics cannot detect.	SME scoring panels; Likert scales; side-by-side comparison with references.
Hallucination Rate	% of outputs containing fabricated, unsupported, or unverifiable claims.	Core risk for GenAI and RAG; key cause of	Human annotation; LLM-as-judge; automated fact-checking; compare output

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

		compliance and operational harm.	claims to ground truth or retrieved sources.
Tool-Call Correctness (PMAS)	Measures whether the agent selects appropriate tools, sequences call properly and interprets responses correctly.	Ensures safety and correctness of multi-step reasoning workflows.	Instrument tool calls; check success rates; simulate tool failures; evaluate sequencing accuracy.
Cascading-Failure Rate	Probability that an upstream failure causes downstream failures in the workflow.	Multi-agent pipelines amplify errors; cascading failures cause systemic risk.	Fault injection: force upstream errors and observe propagation.

Once documentation sufficiency is confirmed, the validator independently verifies performance rather than relying solely on developer-reported results. Unlike traditional ML model validation—which focuses on statistical benchmarking—GenAI systems require creation of "grounded truth" datasets, reference answers, and gold-labelled outputs to objectively measure semantic quality and factual correctness [9-11]. Validators must build or obtain evaluation datasets (e.g., curated questions, policy queries) and compute accuracy, semantic similarity, grounding fidelity, and hallucination rates using both automated metrics and human review [8,21]. For reliability testing, validators develop a structured test plan covering prompt/input variation, N-run consistency, stress scenarios, noise and corruption tests, ambiguity handling, and adversarial perturbations. This ensures independent, reproducible evidence that the system behaves reliably across expected and edge-case conditions.

Table 3: Outcome Reliability Tests

Test	GenAI /RAG	PMAS System	Metric
N-Run Consistency / Stability	Test variation across repeated runs of same prompt	Test stability of each module on repeated structured inputs	Output Variance Score, Semantic Similarity Std-Dev, Stability Index
Prompt / Input Robustness	Paraphrases, synonyms, reordered prompts	Structured input noise, OCR variations, missing fields	Robustness Score, Divergence Rate, Error Sensitivity Index
Ambiguous / Edge Case Handling	Model must handle vague or incomplete prompts safely	PMAS must escalate or fail safely on partial/invalid structured inputs	Safe-Handling Rate, Refusal Accuracy, Escalation Accuracy
End-to-End Outcome Reliability	Output correctness under normal and stressed conditions	Business-outcome correctness for entire pipeline	Task Success Rate, Functional Accuracy, Business-Outcome Score
Noise Robustness	Typos, formatting noise, incomplete instructions	OCR errors, malformed data, corrupted intermediate outputs	Noise Robustness Score, Error Tolerance Index

3.5 Pillar 4 -Risk & Control Assurance:

Pillar 4 evaluates whether the AI system has effective, measurable, and auditable controls that mitigate security, privacy, compliance, and operational risks associated with GenAI, RAG, and multi-component pipelines. Because LLM-based systems generate content dynamically and may integrate with tools, APIs, or enterprise data sources, controls must extend beyond traditional model governance to include prompt safety, data handling protections, guardrails, privacy management, supply-chain assurance, audit logging, human-in-the-loop escalation, and post-processing filters. Validators must assess whether these controls are designed, implemented, monitored, and effective, and whether the model's behaviour remains safe under adversarial, ambiguous, or boundary conditions. This includes evaluating prompt-injection defences, red-team test results, access governance, PII safeguards, fallback logic, override logging, and the completeness of audit trails [14-15,19]. The validator must confirm that all critical steps in the

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

pipeline are logged, including prompting, retrieval, tool calls, agent interactions, safety filter decisions, and post-processing, to enable traceability and incident investigation. Table 4 presents the key risk and control assurance requirements for GenAI systems.

3.6 Pillar 6 - Continuous Monitoring & Lifecycle Management

Pillar 5 ensures that the AI system remains reliable, safe, compliant, and performant throughout its operational life. Because GenAI outputs evolve with prompts, context, data sources, updates to the LLM provider, and changes in organisational documents or policies, continuous monitoring is essential to detect degradation, drift, safety failures, and behavioural shifts. This pillar covers ongoing performance tracking, drift detection, threshold-based alerts, human feedback loops, incident management, and rules for re-validation after material changes. Continuous monitoring combines reference-based evaluation, context-based checks, LLM-as-judge scoring, and human qualitative assessment to provide a multi-layered view of system stability [11, 16, 22]. The objective is to ensure that the model remains aligned with intended use, produces safe and compliant outputs, and adapts appropriately as the environment evolves. Table 5 summarises the GEN-5 monitoring framework.

4. CONCLUSION

As organisations increasingly rely on AI for decision support, automation, compliance operations, and customer-facing interactions, the need for rigorous, transparent, and repeatable validation has never been more critical. Traditional Model Risk Management practices, exposing gaps in how organisations evaluate conceptual soundness, performance, safety, and ongoing reliability. While existing regulatory standards such as SR 11-7, SS1/23, and the 2025 MAS AI Guidelines provide essential foundations, they do not prescribe how to validate systems whose behaviour depends on prompts, retrieval context, tool calls or orchestrated workflows. This paper introduced GEN-5, a five-pillar validation and assurance framework that adapts established MRM principles to the unique risks and characteristics of modern AI systems.GEN-5 provides a structured, policy-aligned approach for assessing AI models across the full lifecycle—from model overview, conceptual soundness, and performance validation to risk and control assurance, and continuous monitoring.

GEN-5 contributes a practical template that fills this gap. It offers MRM practitioners, auditors, and AI governance teams a comprehensive yet operational framework to understand, evaluate, and mitigate the unique risks introduced by Generative AI. By aligning AI validation with regulatory expectations and emerging industry best practices, GEN-5 supports safer deployment, improved resilience, and responsible use of AI across sectors.

Table 4: Control Assurance Measures

Control Area	What Validators Must Check	Why It Matters	Control Metrics
PII Protection & Data Privacy Controls	Verify PII masking, redaction, anonymisation, secure embeddings, domain filtering, RAG metadata restrictions. Ensure PII detectors operate at both input and output layers.	Prevents privacy breaches, regulatory violations, and unintentional model memorisation leakage.	PII Leakage Rate, Redaction Accuracy, Embedding Privacy Score, Metadata Exposure Rate
Prompt Injection & Jailbreak Prevention	Review guardrails (regex filters, policy prompts, safety layers), input sanitisation, adversarial prompt resistance, and boundary constraints. Validate red-team results.	LLMs are vulnerable to malicious prompts that override instructions or extract sensitive information.	Prompt Injection Success Rate, Safety Filter Block Rate, Adversarial Robustness Score
Red-Teaming & Adversarial Testing Controls	Confirm that structured red-team tests are performed for safety, bias, harmful content, jailbreaks, misinformation, toxicity. Review testing frequency and severity scoring.	Identifies vulnerabilities before attackers exploit them. Required by MAS/AI governance frameworks.	Adversarial Attack Success Rate, Severity Index, Patch Coverage Ratio

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Explainability, Transparency, & Reasoning Controls	Check availability of model explanations, rationale summaries, citation correctness (RAG), tool-call traces (agentic), and decision logs.	Required for trust, auditability, regulator review, and model debugging.	Explanation, Completeness Score, Attribution Accuracy, Trace Coverage Rate
Human-in-the-Loop & Escalation Rules	Ensure escalation triggers exist for ambiguity, uncertainty, low confidence, missing retrieval, or inconsistent tool calls. Review HITL workflows.	Reduces automation bias, prevents high-risk actions, and supports override for sensitive use cases.	Escalation Accuracy, HITL Trigger Rate, False-Negative Escalation Rate
Vagueness / Ambiguity Checks	Validate that vague prompts are flagged or clarified, and model does not hallucinate or overcommit when instructions are unclear.	Ambiguous inputs are a major source of hallucination and unsafe answers.	Ambiguity Detection Rate, Safe-Handling Rate, Clarification Prompt Frequency
Access Governance & Role-Based Access Control	Check RBAC for system prompts, model endpoints, vector DB, retrieval domains, and tool integration. Validate token scopes and privilege restrictions.	Prevents misuse, unauthorised data exposure, and privilege escalation.	Access Violation Incidents, RBAC Coverage Rate, Credential Hygiene Score
Override Logging & Contingency Controls	Verify that overrides (developer override, safety bypass, forced output) are logged with user ID, timestamp, reason. Review fallback logic for retrieval failures and tool failures.	Ensures no silent override of controls; crucial for audit and investigations.	Override Frequency, Fallback Success Rate, Unlogged Override Rate
Post-Processing Filters	Validate filters for toxicity, classification, profanity, safety categories, bias mitigation, and policy alignment. Confirm they apply consistently.	Controls the final output layer where unsafe content may escape if upstream checks fail.	Toxicity Filter Accuracy, Content Rejection Rate, False-Negative Safety Rate
Warning & User-Flag Messages	Check if the system warns the user when uncertainty is high, grounding is weak, or domain coverage is limited.	Avoids overconfidence and reduces user misinterpretation of generative outputs.	Warning Frequency, User-Flag Rate, Uncertainty Alignment Score
Audit Logging & Traceability	Ensure complete logs : prompt → retrieval → → LLM generation → tool calls → guardrail decisions → post-processing → final output.	Enables reconstruction of behaviour, forensic analysis, and compliance audits.	Trace Completeness Score, Log Integrity Score, Missing Log Entry Rate
Third-Party Model Risk Controls	Review LLM provider documentation, model card disclosures, update frequency, embedding safety, model hosting risk.	Third-party LLM/API failures or vulnerabilities directly impact safety.	Vendor Compliance Score, Model Drift Score, API Failure Rate

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Table 5: Continuous Monitoring Framework

Monitoring Category	What Is Monitored	Purpose	Key Metrics
Reference-Based Monitoring	Gold-set questions, benchmark tasks, labelled scenarios	Detect accuracy drift and semantic degradation	Accuracy %, Factuality Score, Semantic Similarity, Grounding Score
Context-Based Monitoring	Live user queries, retrieval context, document changes, metadata health	Identify contextual drift and content freshness issues (RAG)	RAG Attribution Score, RAGAs, DeepVal, QuestEval, QAFactEval, ROUGE-C
LLM-as-Judge Monitoring	Automated evaluation of correctness, compliance, safety, grounding	Scalable monitoring between formal validations	LLM-Judge Correctness Score, Safety Score, Attribution Score
Human Feedback Loop (HITL)	Upvotes/downvotes, SME evaluations, flagged outputs	Capture domain-specific errors , compliance gaps	Human Rating Score, Flag Rate, Review-to-Fix Cycle Time
Safety & Guardrail Monitoring	Toxicity filters, PII detection, prompt-injection blocks, escalation triggers	Ensure ongoing safe operation and policy alignment	Toxicity Rate, PII Leakage Rate, Prompt-Injection Block Rate, Escalation Trigger Rate
Performance & Quality Monitoring	Latency, throughput, tool-call success, retrieval errors	Maintain operational quality and reliability	Latency, Timeout Rate, Tool-Call Success %, Error Rate
Threshold-Based Alerting	Breach of accuracy, grounding, safety, latency, or drift thresholds	Trigger investigations and corrective action	Threshold Breach Count, Alert Severity Index
Change Management & Re-Validation	Model version changes, embedding model updates, content re-indexing	Decide when partial/full re-validation is needed	Model Change Log, Re-Validation Trigger Count

REFERENCES

1. U.S. Federal Reserve and Office of the Comptroller of the Currency, Supervisory Guidance on Model Risk Management (SR 11-7), Washington, DC, Apr. 2011.
2. Bank of England and Prudential Regulation Authority, SS1/23 – Model Risk Management Principles for Banks, London, UK, May 2023.
3. Monetary Authority of Singapore, FEAT Principles and Model AI Governance Framework (Ver. 2), Singapore, Jan. 2023.
4. National Institute of Standards and Technology (NIST), AI Risk Management Framework 1.0, Gaithersburg, MD, Jan. 2023.
5. Organisation for Economic Co-operation and Development (OECD), OECD Principles on Artificial Intelligence, Paris, France, 2019.
6. European Commission, Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, Brussels, Belgium, Apr. 2019.
7. International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC), ISO/IEC 42001:2023 – Artificial Intelligence – Management System, Geneva, Switzerland, Dec. 2023.
8. J. Huang, M. Chang, and E. Chi, "A Survey on Hallucination in Large Language Models: Taxonomy and Mitigation ,2309.04843, 2023.
9. J. Yu, S. Wang, and Y. Li, "Evaluating Hallucinations in Retrieval-Augmented Generation Systems," arXiv preprint arXiv:2402.11247, 2024.
10. T. Ribeiro, C. Wang, and K. Zhou, "Faithfulness and Factuality in Language Models: Metrics and Evaluation," Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2023.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 04 April 2026
DOI: 10.47191/ijcsrr/V9-i4-24, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

11. "HalluLens: A Benchmark for Detecting and Measuring Hallucinations in Generative Models," Open-Source Toolkit Documentation, HalluLens Project, 2023.
12. "RAGAs: Retrieval-Augmented Generation Assessment Suite," GitHub Repository, 2024.
13. "eRAG: Evaluation Framework for Retrieval-Augmented Generation Systems," GitHub Repository, 2024.
14. Deloitte LLP, Responsible and Reliable Generative AI: Extending Model Validation and Governance Practices, White Paper, London, UK, 2024.
15. Wells Fargo, Responsible AI and Model Risk Management: Building Trust in Generative AI, Corporate Publication, San Francisco, CA, 2024.
16. Citigroup Inc., Generative AI Risk and Governance Playbook, White Paper, New York, NY, 2024.
17. PricewaterhouseCoopers, Operationalizing Model Validation for Generative AI Systems, Advisory Insight, London, UK, 2023.
18. Barbera, Ai privacy risks & mitigations—large language models (llms), European Data Protection Board,2025
19. OpenAI, "Best Practices for Red-Teaming Large Language Models," Technical Report, OpenAI Inc., 2023.
20. S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," arXiv preprint arXiv:2506.04133 [cs.AI], Sep. 2025
21. S. Chen, Y. Liu, W. Han, W. Zhang, T. Liu, A survey on llm-based multi-agent system: Recent advances and new frontiers in application (2025)
22. D. Biran and C. Cotton, "Explainability and Trust in Autonomous Systems," IEEE Trans. on Human-Machine Systems, vol. 52, no. 5, pp. 801–813, 2022.
23. Monetary Authority of Singapore, *"Consultation Paper on Guidelines on Artificial Intelligence Risk Management (AIRG),"* Singapore, 2025

Cite this Article: Sinha, N. (2026). Adapting the Five Pillars of Model Risk Management for Generative AI: The GEN-5 Validation Framework. International Journal of Current Science Research and Review, 9(4), pp. 1913-1924. DOI: https://doi.org/10.47191/ijcsrr/V9-i4-24