Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Establishing AI State of Control: An Error Taxonomy and Assurance Framework for Transformer-Based Probabilistic Systems in Regulated Life Sciences

Kayla Britt

kayla@brittbiocomputing.com

Britt Biocomputing

Research Article

Keywords:

Posted Date: October 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10476545/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Traditional deterministic risk management frameworks, such as ICH Q9(R1) and classical Failure Mode and Effects Analysis (FMEA), are structurally insufficient for governing transformer-based probabilistic AI (e.g., Large Language Models and Retrieval-Augmented Generation) in GxP environments. Current Quality Management Systems (QMS) lack the vocabulary to classify stochastic, context-dependent AI failures, leading to a gap in deviation logging, root-cause analysis, and CAPA generation.

Methods

This paper introduces a novel Two-Dimensional Probabilistic Error Taxonomy that deconstructs AI failures by mapping six distinct Error Types (e.g., Fabrication, Misinterpretation, Boundary Violation) against six Origination Points (e.g., RAG Layer, Model Inference, Agent Orchestration). To operationalize this taxonomy, the framework integrates an "Irreducibility Gate" into classical FMEA scoring, formally bifurcating AI failure modes into *reducible* errors (addressed via engineered CAPAs) and *irreducible* errors (requiring scope restriction and bounded containment).

Results

By mapping specific failure vectors to distinct risk pathways, the taxonomy serves as a comprehensive diagnostic engine. This risk logic is directly mapped to the FDA's three-stage process validation lifecycle and recent draft credibility guidelines for AI, establishing a structured methodology for bounded variance testing. The utility of the taxonomy is demonstrated through a practical case study of a RAG-augmented LLM used for batch record summarization.

Conclusion

The proposed Two-Dimensional Error Taxonomy bridges the gap between abstract algorithmic risk and practical pharmaceutical quality assurance. It provides the foundational scaffolding required to define a system's context of use, bound its variance, and maintain probabilistic foundation models in a continuous state of control.

1. Introduction: The Limits of Deterministic Risk Management in Probabilistic AI

Classification is critical for identifying error types in GxP settings. ICH Q9 (R1) [1] currently serves as the global standard for risk management implementation in GxP environments. It provides a structured framework for assessing, controlling, reviewing, and documenting risks. Similarly, Failure Mode and Effects Analysis (FMEA) [2] is one of the most commonly used methodologies for structured risk assessment in Computerized System Validation (CSV). Most pharmaceutical and biotechnology organizations with mature quality departments have a Quality Management System (QMS) incorporating these frameworks to recognize and mitigate common CSV failure modes.

However, as probabilistic artificial intelligence (AI) systems become more widespread in GxP settings, there is a need for a structured framework to detect, classify, and control probabilistic-specific errors. While the majority of large pharmaceutical organizations are not yet deploying probabilistic, transformer-based systems at scale [3], tools like Microsoft Copilot and AI-augmented QMS platforms are quietly being introduced into workflows without rigorous procurement due diligence or established governance procedures.

To address this, the ISPE GAMP® AI Guide (led by Brandi Stockton, Eric Staib, and Martin Heitmann) [4] and the Machine Learning Risk & Control Framework (Blumenthal, Erdmann, Heitmann, Lemettinen, & Stockton, 2024) [5] establish the risk-management and control process for AI use cases, building on the foundational GAMP® 5 (Second Edition) guide [6]. The ML Risk & Control Framework [5] enables risk evaluation throughout the full AI system life cycle. It addresses the criticality of risks across various contexts of use, provides a typical risk inventory throughout the ML subsystem lifecycle, and dictates appropriate risk mitigation strategies and lifecycle governance. It maps GxP hazard severity by clustering autonomy/control design levels against potential risk impact, and then outlines a structured process for identifying hazard clusters (e.g., initial data set quality, data split, model design, model training, model evaluation, deployment and release, data quality in operation, and human interaction and monitoring).

Collectively, the existing guidance documents provide organizations with a roadmap for mapping risks throughout the ML lifecycle and corresponding risk mitigation strategies. However, transformer-based probabilistic AI—which, for the purposes of this paper, encompasses text-based Large Language Models (LLMs) as well as Vision-Language Models (VLMs) and multimodal architectures—is inherently stochastic and context-dependent. The same input can produce an output that remains within the validated boundary in one context, while producing an out-of-bounds output in another.

To demonstrate the gap in the existing probabilistic governance guidance, take the following example of a common text-based Large Language Model (LLM) use case. A large pharmaceutical organization deploys an LLM augmented with a Retrieval Augmented Generation (RAG) to pull current, approved documents from the QMS for deviation triage. The operator involved in the manufacturing process enters the deviation into the QMS, after which it is routed to the deviation-triage LLM. The LLM processes the text input and filters through previous quality events to pattern-match to the most likely root cause. The RAG pulls the relevant SOP from the QMS. The deviation is then routed to the quality engineer, who sees the initial operator's report, the LLM's primary, secondary, and tertiary root causes, each with a confidence score.

A robust HITL architecture requires the quality engineer to enter the root cause and justification manually if the validated confidence score is < 90%. The quality engineer must click and verify all referenced quality documents. Upon clicking on the SOP, the quality engineer searches the QMS and is unable to locate the effective document. The actual failure mode? The RAG successfully retrieved the correct record, but the triage LLM ignored the context during generation; instead, it hallucinated a non-existent document that looks correct in structure, but is incorrect in substance.

If a traditional deviation investigation only labels this as a "Software Error," the root cause, a contextual retrieval failure, is completely missed. Therefore, a one-dimensional analysis that classifies errors by type only is structurally insufficient; it obscures the critical information needed to conduct a comprehensive risk assessment, identify the root causes, and map appropriate controls. Existing QMS do not have the vocabulary or practical map to describe what "kind" of error an LLM committed.

In this paper, a novel two-axis classification by error type and origination point is introduced. These intersections map to individual cells with an assigned irreducibility gating score, followed by a structured framework for generating validation assurance evidence. Error classification enables structured risk assessment, deviation logging, and trending, while reducibility gating dictates whether each failure cell maps to an engineered Corrective and Preventive Action (CAPA) or is identified as formally irreducible. Subsequently, evaluation metrics and HITL implementation provide the substrate for initial and ongoing monitoring.

The granular steps for generating specific risk-based controls are outside the scope of this paper. The ISPE GAMP® AI Guide (2025) [4], *Governing Large Language Models in Regulated Pharmaceutical Environments: A Risk Engineering Framework* [7] and *Seven Control Layers for LLMs in GMP Decision-Making* [8] collectively outline orthogonal, complementary frameworks for the creation of a robust control layer strategy. As a whole, the taxonomy proposed here provides the comprehensive diagnostic engine required for both governing probabilistic failures and generating assurance evidence. This approach aligns closely with established literature and ICH Q8 [9] regulatory precedent: analogous to pharmaceutical process validation, we define the design space and characterize behavior at its edges. This paradigm assures bounded risk for deep learning foundation models, including Large Language Models (LLMs), Vision-Language Models (VLMs), Retrieval-Augmented Generation (RAG) architectures, and multi-agent orchestrations.

This paper makes three primary contributions to the field of probabilistic AI governance in GxP settings:

1. It introduces a novel Two-Dimensional Probabilistic Error Taxonomy, classifying failures by both error type and origination point.
2. It integrates the concept of irreducibility gating into a structured framework for classifying failure modes, whereby errors can be mapped to individual cells, bifurcating into either an engineered CAPA (reducible failures) or scope restriction (irreducible failures).
3. It maps this risk logic to the FDA's draft Risk-Based Credibility Assessment Framework (*Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products*, January 2025 [10]).

The remainder of the paper outlines both dimensions of the two-axis taxonomy: error type and origin point. Each axis breaks down into six broad classes, each of which encompasses several subclasses. A six by six matrix with a total of thirty-six (36) cells allows practitioners to identify the precise cell matching the known failure mode. A structured process for integrating the irreducibility gate follows, along with a proposed three-stage methodology for validation assurance generation mapped to the FDA's three-stage process validation lifecycle. Finally, the paper concludes with an example that demonstrates the methodology in practice.

2. The Two-Dimensional Error Taxonomy: Deconstructing Probabilistic Failures

Per ML Risk & Control Framework [5], after the quality risk management process has been initiated, and the hazard severity of the potential GxP hazards identified has been established, the risk assessment is initiated. The framework maps hazards to the eight machine learning (ML) lifecycle phases established in the GAMP® 5 2nd Edition guide.

Georgi Tancev (2026) introduced an interoperating framework in the paper "Governing Large Language Models in Regulated Pharmaceutical Environments: A Risk Engineering Framework" [7]. The framework includes a failure taxonomy organized around three deployment architectures: pure LLM, retrieval-augmented generation, and LLM-based agents, and a governance framework including a Layered Control Architecture, LLM Failure Mode and Effects Analysis (LLM-FMEA) with an irreducibility flag, a LLM Deployment Integrity Level (LDIL) framework.

This paper builds on both the ML Risk & Control Framework [5] and Tancev's governance frameworks (Layered Control Architecture, LLM-FMEA with an irreducibility flag, and a LLM Deployment Integrity Level (LDIL) framework [7]). The two-dimensional probabilistic error taxonomy is the complementary framework to the ML Risk & Control Framework and the governance frameworks; it maps errors to an appropriate control and lifecycle phase.

Where Tancev's governance frameworks outline the layered control architecture to bound the errors, the LLM-FMEA computes a Risk Priority Number to assess the overall risk of each error, and the LDIL maps errors by GxP impact, the probabilistic error taxonomy provides the vocabulary for both classifying errors and identifying the root cause.

The two-dimensional taxonomy is largely orthogonal to Tancev's failure taxonomy; where the failure taxonomy classifies by irreducibility, the two-dimensional taxonomy classifies by error type and origin. A granular breakdown of all subclasses is provided in Table 1.

2.1 Dimension 1 (Error Type): To classify the exact nature of the probabilistic failure, six primary error classes are established.

- Class 1: Fabrication (e.g., inventing citations or data points).
- Class 2: Misinterpretation (e.g., reasoning errors or omissions).

- Class 3: Contextual Misapplication (e.g., context mismatch or specification non-compliance).
- Class 4: Confidence (e.g., overconfidence or under-confidence).
- Class 5: Boundary Violation (e.g., scope creepy or authority creep).
- Class 6: Population Bias (e.g., demographic subgroup bias or pharmacovigilance/safety-signal bias).

2.2 Dimension 2 (Origin Point): To trace the failure back to its architectural root cause, six origination points are defined:

- Origin 1: Training Data (e.g., data poisoning or data privacy).
- Origin 2: Retrieval/RAG Layer (e.g., chunking failures or embedding drift).
- Origin 3: Model Inference (e.g., hallucination or sycophancy).
- Origin 4: Human-AI Interface (e.g., automation bias or confidence miscommunication).
- Origin 5: Agent Orchestration (e.g., quality and state degradation across iterations or compounding propagation - where multiple errors interact across stages).
- Origin 6: Supplier (e.g., silent model updates or deprecation).

The diagram depicted in Fig. 1 demonstrates the dual axes of error types and origination points, each with a sub-classification(s). This enables organizations to log deviations, investigate, and generate a defined CAPA, under ICH Q10 [11]. Each deviation may map to vectors.

Both Class 4 (Calibration) and Class 6 (Bias) are novel failure classes not previously established in GxP probabilistic literature.

3. The Irreducibility Gate: Adapting FMEA for Non-Deterministic Systems

3.1 The Irreducibility FMEA Addition

Tancev [7] introduces the concept of formally distinguishing failure modes between remediable (reducible by engineering effort) and irreducible (cannot be reduced to zero through known engineering approaches): the "irreducibility partition." The paper articulates the unique nature of probabilistic failures originating from Application Programming Interface (API)-accessed LLM deployments; "undesired outputs are suppressed probabilistically rather than eliminated categorically." Therefore, a tailored risk logic is required to surface failure modes that may exist structurally but go undetected.

He proposes incorporating irreducibility (I) into the classical FMEA calculation. This computes a Risk Priority Number $RPN = S \times O \times D \times I$ from four total parameters: Severity (S), Occurrence (O), Detectability (D), and Irreducibility (I). The two-dimensional probabilistic error taxonomy provides the complementary framework to map errors to a respective RPN. This paper proposes the novel mapping of each cell from the Two-Dimensional Error Taxonomy matrix in Fig. 1 to Irreducibility (I).

The irreducibility gate establishes a differentiated control approach for deciding the type of control evidence required for each occupied cell. This enables the execution of an operational FMEA scoring exercise, where each error, mapped by Error Type and Origin, is assigned a Risk Priority Number (RPN). Irreducible cells get detection-coverage + scope criteria with occurrence-rate targets explicitly suspended; reducible cells get engineered-down criteria.

Tancev (2026) [7] notes that the majority of the impossibility results referenced in his paper (Xu et al. [12], Moshin et al. [13], and Nourizadeh [14]) have not been fully peer-reviewed. Only Asher et al. [15] has been fully peer-reviewed. This means that the empirical basis for the irreducibility determination of failure modes must be interpreted in this context.

3.2 Strategic Divergence: Engineered CAPAs vs. Scope Restriction

The irreducibility gate ensures that failure modes that cannot be engineered out completely are explicitly flagged in the Quality Management System (QMS). For these occurrences, root cause analysis will inherently trace back to the mathematical structure of the system itself, rather than a preventable operational event. Consequently, these failure modes require documented scope restriction as an additive risk management effort.

For irreducible failures, the occurrence (O) parameter in the Risk Priority Number $(RPN = S * O * D * I)$ is fixed. Therefore, containment efforts must focus on reducing severity (S) by limiting deployment scope to use cases where irreducible failure is bounded. Alternatively, organizations must accept the residual risk with explicit failure boundary documentation and additional architectural containment, such as HITL cognitive forcing functions or seeded quality assurance checks.

In this context, a traditional aggregate RPN obscures meaningful information about the risk profile; therefore, each failure mode mapped to a cell in the two-dimensional error taxonomy must be formally designated by its irreducibility status.

The Two-Dimensional Error Taxonomy is the central focus of this paper, as it is foundational to each phase of governance. It serves as the scaffolding for the initial risk assessment with irreducibility gating, and for computation of the FMEA. These risk scores guide validation execution planning (Section 4). Retrospectively, the Two-Dimensional Error Taxonomy gives practitioners the vocabulary for deviation logging, classification, and investigation (Section 5).

4. The GxP Probabilistic Validation Lifecycle: Adapting FDA Guidelines for AI

Once the risk management process is complete, the next step is generating appropriate validation assurance evidence for the context-of-use (COU) in a risk-based manner, based on the risk scores and irreducibility gate mapping generated during the risk-assessment process.

Here, a structured process for evidence-generation is proposed.

The evaluation design stages are mapped to the existing three-stage process validation lifecycle established in the 2011 FDA guidance document “Process Validation: General Principles and Practices” [16] (See Fig. 4). Each stage incorporates steps from the FDA’s 2025 draft credibility assessment document “Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products” [10].

4.1 Stage 1: Credibility Planning and Boundary Setting (Aligns with FDA PV Stage 1 - Process Design)

4.1.1 Step 4 of FDA Credibility Assessment: Establish Credibility Goals

This step involves defining acceptable probabilistic bounds within which the model’s inputs and outputs remain in an assured state. For instance: “The system must exhibit zero Class 1 (Fabrication) errors at Origin 2 (Retrieval), and maintain variance within 5% across N equivalent prompts.”

4.2 Stage 2: Bounded Variance Testing (Aligns with FDA PV Stage 2 - Process Qualification)

4.2.1 Step 5 of the FDA Credibility Assessment: Execute the Plan

At this stage, characterization of the stochastic envelope measures how far a model's output moves across contextually equivalent, non-identical inputs, bounded against pre-specified thresholds. This converts an unreproducible output into a reproducible distribution of outputs: the form of reproducibility that probabilistic validation actually requires. This corresponds to Stage 2 (process qualification).

4.3 Stage 3: Continuous Monitoring and HITL (Aligns with FDA PV Stage 3 - Continued Process Verification)

4.3.1 Step 6 and 7 of the FDA Credibility Assessment: Document Results and Determine Adequacy

During the final stage of the GxP probabilistic validation lifecycle, the results from bounded variance testing in stage 2 are documented, and a credibility assessment is generated. Additional controls, including human-in-the-loop (HITL) requirements and human performance qualification, are implemented where necessary to mitigate risk. Finally, the sponsor determines whether the AI model is suitable for the defined context-of-use within the system’s boundaries prior to deployment. Crucially, this continued process verification must utilize the Two-Dimensional Error Taxonomy to log and trend any deviations (e.g., Origin 6a: Silent model updates or Origin 2b: Embedding drift) that emerge post-deployment, ensuring the system remains in a state of control.

5. Operationalizing the Framework: A Case Study in Batch Record Summarization

A pharma company is piloting a RAG-augmented, locally hosted LLM to summarize manufacturing batch records and identify all critical process parameters (CPPs) for the Quality Assurance (QA) lot release team.

5.1 Prospective Risk Assessment and Irreducibility Mapping

The organization’s QA team first works cross-functionally with the IT and Manufacturing departments to map out the architecture of the system. This initial process of mapping the system’s architectural elements directs which columns of the Two-Dimensional Error Taxonomy are assessed during the risk analysis. The cross-functional team utilizes the 6 x 6 Two-Dimensional Error Taxonomy to identify the potential failure modes of the LLM context-of-use:

- Failure Mode 1: Origin 2 (Retrieval/RAG) x Class 2 (Misinterpretation), mapping to chunking failure.
- Failure Mode 2: Origin 2 (Retrieval/RAG) x Class 1 (Fabrication), mapping to grounded hallucination.
- Failure Mode 3: Origin 1 (Training Data) x Class 1 (Fabrication) mapping to poisoned source data.
- Failure Mode 4: Origin 1 (Training Data) x Class 2 (Misinterpretation) mapping to omission in source.
- Failure Mode 5: Origin 4 (Human-AI Interface) x Class 1 (Fabrication) mapping to unverified output.
- Failure Mode 6: Origin 4 (Human-AI Interface) x Class 2 (Misinterpretation) mapping to user misinterpretation of the output.

The team then calculates the FMEA RPN score and irreducibility gate for each identified risk. Failure Modes 1 and 2 are classified as irreducible, given the architectural constraint that requires the model to parse the batch record in chunks and the model’s fundamental probabilistic processing nature. The system is bound by strict system prompts instructing it to only use retrieved text. The baseline risk tier is established as High Influence / Moderate Consequence.

5.2 Process Qualification: Probabilistic GxP Validation Testing

The Quality/Validation team then works jointly with the IT/Data Science team and Subject Matter Experts (SMEs) to execute process qualification for the LLM.

The qualification strategy shifts from point-in-time testing to bounded variance testing, to ensure the model’s outputs fall within the validated range. With Quality/Validation oversight, the IT/Data Science team runs the approved prompts for historical batch records (N = 50) through the LLM. The SMEs evaluate and score the responses against the SME “ground truth”. The team drafts the execution protocol with acceptance criteria based on the identified failure modes and their FMEA scores/irreducibility gating. For instance, given that failure modes 1 and 2 are identified as high-risk and irreducible, acceptance

criteria are structured to address the additional risks: the model must score 50/50 on critical process parameter identification during bounded variance testing and the human-AI interface must be qualified.

During Stage 3 Continuous Monitoring, a QA reviewer spots an error in the LLM's summary. The LLM correctly flagged a temperature excursion (e.g., the reactor reached 180°C), but it hallucinated the root cause, stating that "a valve malfunctioned." The batch record explicitly stated the root cause was "operator input error." A traditional QMS would classify this broadly as a "Software Error" or "AI Hallucination," leaving the validation engineer without a defined path for root-cause analysis or a CAPA.

5.3 Post-Deployment Deviation Management and CAPA Routing

- **Dimension 1 (Error Type)**: Using the two-dimensional error taxonomy, QA maps this to **Class 1d (Rationalization fabrication)**: inventing false logic for a correct data point.
- **Dimension 2 (Origin)**: The QA team investigates the RAG logs. The vector database successfully retrieved the exact sentence stating "operator input error." Therefore, the failure did not originate in the RAG layer (Origin 2). The model simply ignored the retrieved context and generated a highly probable, but false, manufacturing reason from its parametric memory. This is mapped to **Origin 3a (Intrinsic Hallucination)**.
- **The Resulting Vector**: Class 1 × Origin 3.
- **Corrective Action (CAPA)**: The QA team creates a Corrective Action in the QMS requiring an architectural control (semantic grounding evaluator) within the validation harness. This “forced check” verifies whether the model’s generated text matches the facts present in the retrieved RAG vector. If a hallucinated output is detected, the evaluator blocks it from being presented for review.

6. Limitations and Future Directions: A Living Framework

This taxonomy is proposed as a foundational, living framework for governing probabilistic errors in GxP settings. It is designed to evolve as empirical case studies accumulate and to accommodate rapidly shifting probabilistic technologies and regulatory expectations. The six error types and origination points provide a durable structure for the majority of regulated life science workflows. However, the Irreducibility Gate (I) is inherited from Tancev’s preprint, meaning its practical integration remains contingent on the peer-reviewed validation of those foundational mathematical proofs. Furthermore, practical implementation and sub-category refinement are expected to iterate the taxonomy over time.

Most notably, elements of the framework were recently adapted and submitted in a formal comment [17] in response to the FDA's AI-Enabled Early-Phase Real-Time Clinical Trial Pilot docket (FDA-2026-N-4390). In that submission, the framework was applied to bridge the gap between abstract algorithmic risk and practical quality assurance/GAMP® guidance by explicitly mapping probabilistic testing assurance directly to the traditional process validation lifecycle.

Emerging global regulatory paradigms are increasingly bifurcating AI applications into two distinct categories: "critical" systems requiring high determinism, and "non-critical" or supportive systems where probabilistic models (e.g., Large Language Models) may be deployed with appropriate oversight. This evolving regulatory stance underscores a gap in current industry practice. To successfully deploy probabilistic AI in supportive GxP workflows, organizations require a standardized methodology to define a system's Context of Use (CoU), mathematically bound its variance, and empirically prove the effectiveness of Human-in-the-Loop (HITL) controls. This paper proposes foundational scaffolding to address this challenge.

6.1 Limitations

The immediate limitation of this matrix lies in the necessity for empirical case accumulation. While the framework categorizes failures observed in current multi-agent and RAG architectures, vector mappings—particularly those intersecting Agent Orchestration (Origin 5) and Supplier updates (Origin 6)—will require validation against longitudinal deviation data from live GxP environments.

As computational methodologies advance, the classification of a failure mode as formally "irreducible" must be periodically reassessed against the current state of engineering capabilities. Finally, while Class 4 (Confidence Miscalibration) and Class 6 (Population Bias) are introduced here as critical dimensions, they demand deeper, dedicated treatment regarding statistical thresholding in future literature.

6.2 Future Directions: The Unstructured Data Imperative

The most urgent frontier for this framework is its application to unstructured data pipelines, particularly in the context of the FDA’s Real-Time Clinical Trials (RTCT) Pilot Program announced in April 2026. The RTCT initiative fundamentally shifts the regulatory paradigm by allowing sponsors to transmit pre-defined safety and efficacy signals to the agency in near real-time, eliminating traditional trial hiatuses.

Achieving this continuous transmission requires the automated extraction of signals from unstructured clinical sources, such as physician notes, pathology reports, and free-text patient-reported outcomes, often via LLM intervention. In this architecture, the traditional manual data-cleaning buffer is removed. An AI error in unstructured extraction (e.g., an omission [Class 2c] originating from a chunking failure [Origin 2a]) results in the immediate transmission of corrupted or incomplete safety signals directly to the regulator.

Applying this Two-Dimensional Error Taxonomy to govern the automated extraction of unstructured Electronic Health Record (EHR) data for live signal transmission represents a critical next step. Future research must adapt the bounded variance testing and continuous monitoring principles established here to secure the integrity of real-time clinical trial data.

Declarations

Author Contribution

K.B. wrote the entire manuscript text.

References

1. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised guideline: quality risk management Q9(R1). Geneva: ICH; 2023.
2. International Electrotechnical Commission (IEC). IEC 60812: failure modes and effects analysis (FMEA and FMECA). 3rd ed. Geneva: IEC; 2018.
3. McKinsey & Company. Scaling gen AI in the life sciences industry. Chicago (IL): McKinsey & Company; 2025 Jan. Available from: https://www.mckinsey.com/industries/life-sciences/our-insights/scaling-gen-ai-in-the-life-sciences-industry.
4. Stockton B, Staib E, Heitmann M. ISPE GAMP AI guide. North Bethesda (MD): International Society for Pharmaceutical Engineering (ISPE); 2025.
5. Blumenthal M, Erdmann C, Heitmann M, Lemettinen M, Stockton B. Machine learning risk & control framework. Pharm Eng. 2024;44(Jan-Feb):12-22.
6. International Society for Pharmaceutical Engineering (ISPE). GAMP 5: a risk-based approach to compliant GxP computerized systems. 2nd ed. North Bethesda (MD): ISPE; 2022.
7. Tancev G. Governing large language models in regulated pharmaceutical environments: a risk engineering framework. arXiv [Preprint]. 2026. Available from: https://www.researchgate.net/publication/404873511_Governing_Large_Language_Models_in_Regulated_Pharmaceutical_Environments_A_Risk_Engine
8. Bechmann V, Albrechtsen P. Seven control layers for LLMs in GMP decision-making. Pharm Eng. 2026;46(Jan-Feb):21-26.
9. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised tripartite guideline: pharmaceutical development Q8(R2). Geneva: ICH; 2009.
10. Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. Draft Guidance. Silver Spring (MD): U.S. Department of Health and Human Services; 2025 Jan.
11. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised tripartite guideline: pharmaceutical quality system Q10. Geneva: ICH; 2008.
12. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. arXiv [Preprint]. 2024. Available from: https://arxiv.org/abs/2401.11817.
13. Mohsin MA, Umer M, Bilal A, et al. On the fundamental limits of LLMs at scale. arXiv [Preprint]. 2025. Available from: https://arxiv.org/abs/2511.12869
14. Nourizadeh M. No red lines: the impossibility of formal safety guarantees in LLMs. PhilArchive [Preprint]. 2026. Identifier: NOUNRL v2.0.
15. Asher N, Bhar S, Chaturvedi A, Hunter J, Paul S. Limits for learning with language models. In: Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). Association for Computational Linguistics; 2023. p. 236-248. doi: 10.18653/v1/2023.starsem-1.22.
16. Food and Drug Administration. Process validation: general principles and practices. Silver Spring (MD): U.S. Department of Health and Human Services; 2011.
17. Britt K, Canterbury J, Sharma T. Comment submitted to U.S. Food and Drug Administration Docket No. FDA-2026-N-4390. Silver Spring: Regulations.gov; 2026.

Table 1

Table 1 is not available with this version.

Figures

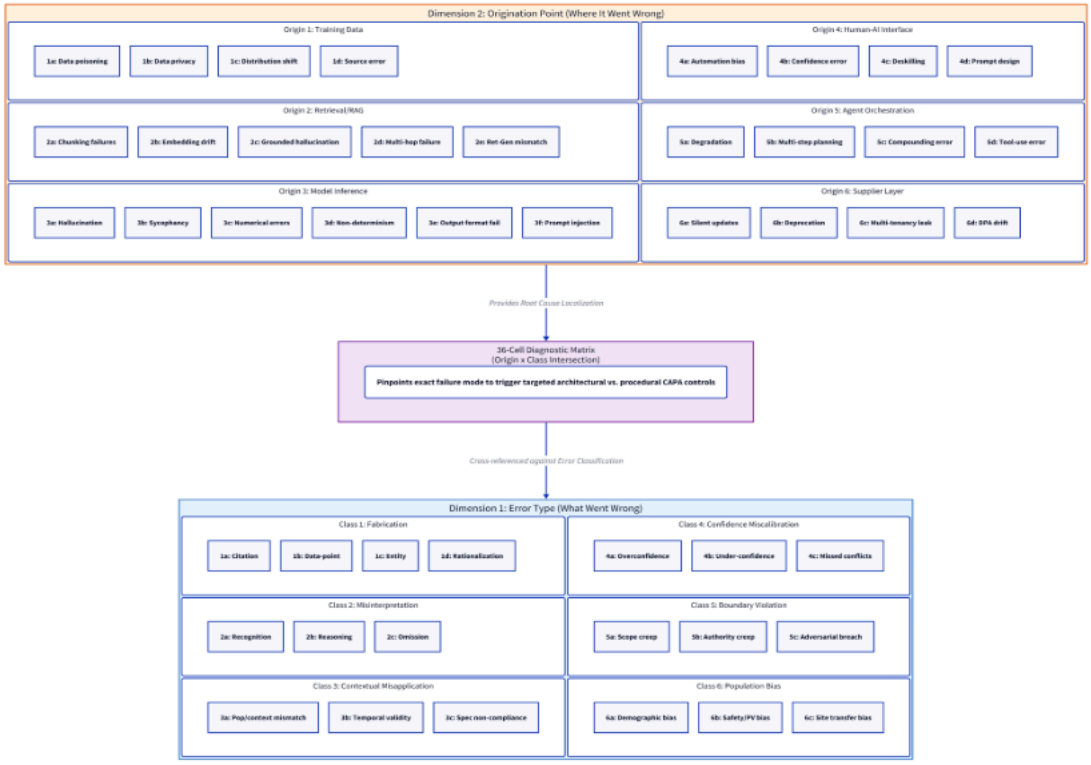


Figure 1

Two-Dimensional Probabilistic Error Taxonomy Table



Figure 2

Risk Assessment Flowchart

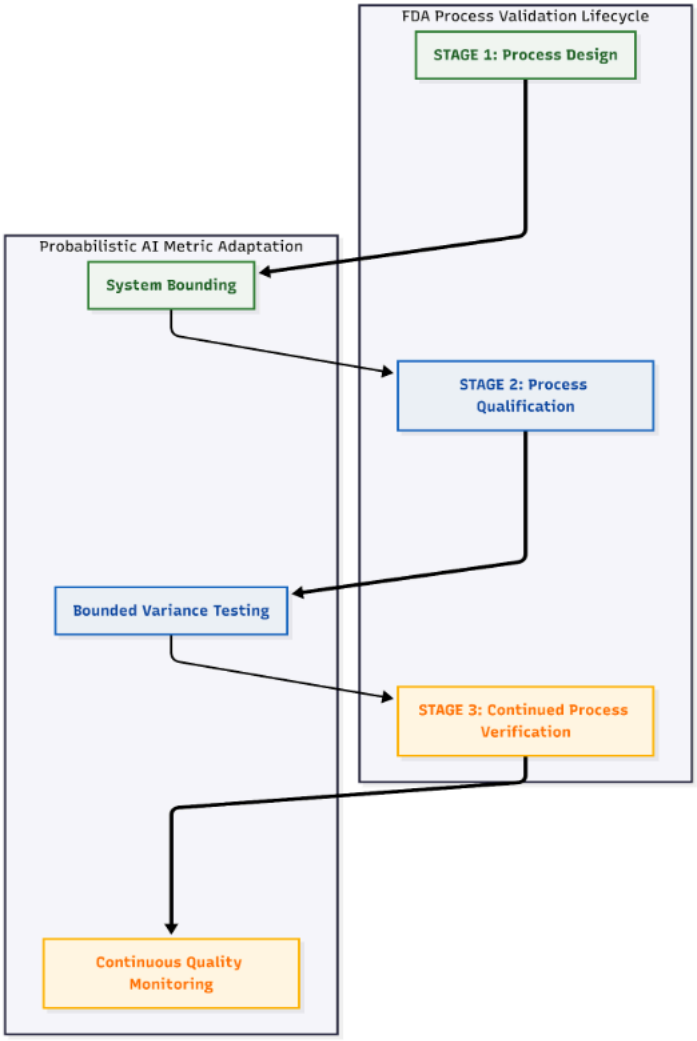


Figure 3

Three-stage evaluation mapped to the FDA's 2011 process validation lifecycle