Changing the Engine While Flying

Maintaining the Validated State When AI Models Change in Pharmaceutical GxP Systems

Reinhold Schilling

Schilling Vigilance White Paper Series

No. 7 | Revision 0 | 25 September 2026

Executive Summary

A pharmaceutical company may validate an AI-enabled system for a defined task, only to find that its supplier subsequently updates the underlying model. The change may be visible and scheduled, or it may occur behind an unchanged application interface. Model weights are only part of the problem: prompts, retrieval sources, embeddings, guardrails, APIs and human workflows can also change the behaviour of the system that was originally validated.

A model change is a change-control trigger, not automatically a full-revalidation trigger. Established GxP principles require companies to control relevant changes and maintain systems fit for their intended use; they do not prescribe complete repetition of initial validation after every AI update. The appropriate response depends on what changed, how much the system relies on the changed component, the consequences of error, and whether existing evidence still applies. [1–4]

This paper proposes a practical, four-class change framework and a decision process for selecting proportionate assurance activities. A non-impacting change may require a documented assessment but no new test. A bounded change may call for focused regression testing. A material performance-affecting change, including many model replacements, will ordinarily need partial revalidation against predefined requirements and representative reference data. A fundamental change may require a new validation baseline.

The central object of assurance is the complete AI-enabled system in its context of use, not the model alone. Its configuration includes the model and its version, instructions, data sources, interfaces, surrounding software, operating procedures and human review. For probabilistic systems, continued fitness may need to be demonstrated through performance boundaries and critical-error analysis rather than identical outputs on every run. [4]

Initial validation remains indispensable. It establishes the approved intended use, risks, acceptance criteria and baseline evidence. Lifecycle assurance then preserves that state through controlled configuration, change detection, impact assessment, proportionate revalidation, approved release and performance monitoring. For externally hosted AI, this is defensible only if supplier arrangements provide sufficient visibility and control to identify the effective production configuration and investigate regulated outputs.

Core proposition: Identify the change. Assess its GxP impact. Test what could matter. Explain what prior evidence still applies. Approve the new configuration. Monitor whether it continues to perform within justified boundaries.

The Hidden Validation Problem

Consider a pharmacovigilance application that extracts case information from source documents. The company validated the application using Model A, approved its intended use and trained case processors to review its outputs. Months later, the supplier substitutes Model B. The application looks the same. Its forms, buttons and workflow appear unchanged.

The visible interface does not establish that the validated system is unchanged. Model B might identify more adverse events but miss a particular seriousness criterion. It might produce better narratives while introducing fabricated dates. It might return different structured fields, refuse certain requests, process long documents differently or increase the time reviewers need to detect errors. A change to the model can therefore affect both the automated result and the effectiveness of the surrounding human control.

Nor is a model swap necessarily a new system requiring the company to discard every item of prior validation evidence. Access controls, an unchanged interface and established operating procedures may remain adequately supported. The task is to determine which parts of the original assurance case still hold and which must be renewed. Risk-based lifecycle management, change control and periodic assessment are established features of pharmaceutical computerised-system governance. [1–3]

This distinction matters commercially as well as technically. Freezing AI indefinitely may be impossible when suppliers retire model versions or address security defects. Repeating an entire initial validation after every bounded update can consume resources without improving control of the newly introduced risk. Neither extreme provides a satisfactory answer to the inspector’s question: Why was this configuration acceptable for this GxP use on the day it produced the regulated output?

What Is Actually Changing?

For change-control purposes, a relevant AI change is any alteration capable of affecting a GxP requirement, output, record, decision, risk control or the ability to reconstruct operation. It need not alter model weights.

A model-version update, replacement with another foundation-model family, retraining or fine-tuning may change output quality and failure patterns. But a revised system prompt can also change what a model prioritises or omits. Retrieval-augmented generation (RAG) can change when its knowledge sources, indexing, chunking, embedding model or ranking logic change. An unchanged generator may then receive materially different evidence. Inference parameters can change variability or cause truncation; guardrails can block valid outputs as well as prevent unsafe ones.

Changes outside the apparent AI component also matter. An API response can acquire a new field or default. A parser can silently drop a value. A revised user interface can make an incorrect suggestion appear more credible. Removing a second reviewer can alter the chance that an error reaches a regulated record. Expanding the application from drafting support to autonomous classification changes its decision role even if no code changes. EMA’s AI reflection paper addresses fitness for purpose across algorithms, models, datasets and processing pipelines, with risk tied to context of use and the model’s influence on decisions. [4]

The validated boundary should therefore follow the GxP function. It should include the model, instructions and parameters; relevant training or fine-tuning arrangements; retrieval and knowledge sources; interfaces and downstream rules; the data and records created; and the people and procedures relied upon to detect or correct errors. This is an assurance boundary, not a claim that each component always needs the same depth of testing.

The boundary also makes cumulative change visible. A prompt adjustment, a retrieval-index rebuild and a change to reviewer workload might each look modest in isolation. Together they may invalidate the assumptions under which performance and human oversight were previously accepted.

The Regulatory Foundation

Regulatory sources support a risk-based answer, but they do not provide a universal pharmaceutical rule for deciding that a particular AI update requires a specified number of tests. It is important to distinguish operative GxP expectations, non-binding guidance, consultation drafts and the practical approach proposed in this paper.

[REGULATORY REQUIREMENT] Current EU GMP Annex 11. The January 2011 revision remains the operative Annex 11 listed in EudraLex Volume 4 at this paper’s cut-off. It applies to computerised systems used in GMP activities. Its provisions address lifecycle risk management, validation of applications, supplier oversight, change and configuration management, and periodic evaluation of whether a system remains in a valid state.

These principles establish the need to control a relevant model change; they do not state that every such change demands complete revalidation. [1]

[REGULATORY GUIDANCE] ICH Q9(R1) and ICH Q10. ICH Q9(R1) frames quality risk management around assessment, control, communication and review, with formality proportionate to uncertainty, importance and complexity. ICH Q10 places change management within the pharmaceutical quality system and calls for evaluation of changes using quality risk management principles. Together, they support documented, risk-proportionate decisions rather than an automatic response based on the technical label of a change. [2,3]

[REGULATORY GUIDANCE] EMA's AI reflection paper. This final reflection paper, adopted in 2024, addresses AI throughout the medicinal-product lifecycle. It connects evidence to intended use, model influence and potential patient or regulatory impact. It discusses suitable data, predefined performance thresholds, monitoring for drift or degradation, and re-evaluation after non-trivial changes in higher-impact applications. It is especially useful for interpreting what fitness for purpose means for AI, but it does not create a single cross-GxP revalidation algorithm. [4]

[REGULATORY GUIDANCE] Domain-specific sources. ICH E6(R3) applies risk-proportionate thinking to clinical-trial systems and the importance of the data and activities they support. EMA's clinical-trial computerised-systems guideline addresses validation, interfaces, change control and lifecycle operation. OECD guidance for GLP computerised systems likewise addresses controlled changes and continuing suitability. In pharmacovigilance, EMA's reflection paper places responsibility on marketing authorisation holders for validating, monitoring and documenting AI performance; the final CIOMS Working Group XIV report provides a non-binding, PV-specific framework for responsible use. [4–8]

[CONSULTATION DRAFT, NOT A CURRENT REQUIREMENT] Revised Annex 11 and proposed Annex 22. The European Commission opened consultation on both texts in July 2025. The revised Annex 11 draft develops lifecycle, supplier and risk-scaled validation expectations. Proposed Annex 22 adds AI-specific material on intended use, test data, acceptance criteria, change assessment and operational monitoring for specified critical GMP manufacturing applications. Neither consultation text should be cited as if it had replaced the operative 2011 Annex 11 or established binding rules for all GxP domains. In particular, the proposed Annex 22's treatment of probabilistic models and critical GMP use of generative AI is a draft manufacturing position, not a general prohibition across pharmacovigilance, clinical research or other pharmaceutical operations. [9,10]

[REGULATORY GUIDANCE WITH LIMITED SCOPE] FDA Computer Software Assurance. As of September 2026, the relevant final FDA document is Computer Software Assurance for Production and Quality Management System Software, issued 3 February 2026. It superseded the September 2025 guidance cited in some earlier research materials. It recommends a risk-based approach to establishing confidence in software used in medical-device production and quality management systems. Its intended-use and proportionate-evidence principles are useful by analogy; the document should not be presented as generally applicable drug-GMP, GCP or GVP guidance. [11]

[DRAFT REGULATORY GUIDANCE] FDA's drug AI credibility framework. FDA's January 2025 draft guidance on AI supporting regulatory decision-making for drug and biological products proposes defining a question and context of use, assessing model risk from influence and decision consequence, and planning and evaluating credibility evidence. Its scope concerns information submitted to support regulatory decisions. It is not a final rule for operational GxP model changes, although its context-of-use reasoning is valuable. [12]

[INDUSTRY GOOD PRACTICE] GAMP and other AI frameworks. ISPE's GAMP 5 Second Edition offers a risk-based computerised-system lifecycle approach. CIOMS Working Group XIV offers PV-specific principles. Voluntary AI risk frameworks can help structure testing and monitoring. These sources inform an effective quality-system design; they are not substitutes for applicable regulations or for a company's documented assessment of its own use case. [8,13,14]

[SEPARATE LEGAL REGIME AND TRANSFERABLE CONCEPT] EU AI Act and medical-device PCCPs. The EU AI Act may impose obligations when a particular system and operator fall within its scope. Its treatment of modifications cannot simply be substituted for a GxP validation decision. Likewise, FDA's predetermined change control plan, or PCCP, approach was developed for specified AI-enabled medical devices. Its useful principle is prospective control of foreseeable changes and their assessment methodology. It is not a general

pharmaceutical GxP requirement. [15,16]

The governing conclusion is deliberately narrower than either “every update requires full revalidation” or “passing a benchmark is enough.” A regulated company must be able to explain how it assessed the change, why it selected its validation scope, and why the resulting system remained fit for the approved use.

From Validation to Lifecycle Assurance

Traditional computerised-system validation often uses a straightforward question: did the implemented function produce the expected result under defined conditions? That question remains essential for deterministic interfaces, calculations, permissions, audit trails and record handling. It is less complete for a probabilistic model whose answer may vary between runs or whose aggregate performance can conceal consequential errors.

The complementary AI question is: Does the complete system remain demonstrably fit for its intended use within predefined performance and risk boundaries? The question does not relax validation. It requires a more precise statement of the task, the operating population, the model’s decision role and the errors that cannot be accepted. EMA’s reflection paper and FDA’s draft credibility framework both place context of use at the centre of the evidence assessment. [4,12]

Intended use describes what the system is authorised to do. Context of use gives that statement operational meaning: which source documents or process data it handles; which populations, products, sites or languages it encounters; who acts on its output; what other controls intervene; and what happens when it is wrong. A document-drafting assistant subject to complete qualified review presents a different risk from a classifier that automatically suppresses cases from a safety queue.

A governed reference dataset turns intended use into testable evidence. It should represent routine inputs and important edge cases, contain qualified reference labels, preserve its provenance and remain sufficiently independent of model development. A locked regression set supports comparison across versions. Fresh challenge cases are also necessary when a change creates plausible new failure modes or repeated testing has made the old set too familiar. The design of these datasets is part of validation, not merely a data-science convenience. [4,8]

Acceptance criteria must be specified before formal testing. Overall accuracy may be appropriate for some tasks but misleading for others. In ICSR intake, a false negative that causes a potentially reportable case to be missed may matter more than several false positives that receive human review. In visual inspection, the relevant question may be sensitivity for particular critical defects. For generative outputs, factual grounding, omitted information, invented statements and invalid structured responses may each need a separate measure. Performance by language, site, document type or other relevant subgroup may reveal a regression that an aggregate score hides.

For a probabilistic model, literal output identity should not be presumed. A test strategy may need repeated runs on a defined subset, analysis of disagreement on critical fields, confidence intervals around estimated performance, and prespecified rules for indeterminate outputs. The aim is not to prove that every future response will be identical. It is to show, with uncertainty made visible, that the system operates within acceptable boundaries for the approved task.

Human review must be analysed with the same seriousness as automated performance. A reviewer may catch model errors only if the source is available, the interface makes discrepancies visible, time is sufficient and the reviewer knows what to verify. Review that exists on an SOP but is ineffective in practice cannot be credited as a strong mitigation. Where a model change alters output style, apparent confidence or workload, evaluate the human–machine workflow, not merely the model in isolation.

Lifecycle assurance begins after a sound initial validation. It requires a controlled production configuration, detection of supplier and internal changes, monitoring against predefined thresholds, investigation of adverse trends and periodic reassessment. Monitoring can detect drift in incoming documents or model performance; it cannot retrospectively replace evidence that should have been obtained before releasing a materially changed configuration. EMA’s reflection paper addresses both predeployment evidence and operational

performance monitoring. [4]

A Practical Change Classification

[AUTHOR'S PROPOSED APPROACH] The following four classes are a governance tool, not a classification mandated by a regulator. They describe the likely assurance response, not the technological sophistication of the update.

Classification should consider intended-use change, model influence, consequence of error, detectability, strength of human verification, novelty and uncertainty, affected requirements, new failure modes, supplier transparency, ability to identify the production configuration, reversibility and cumulative changes. A change described by a supplier as "minor" may still be material when it affects a critical field. Conversely, a substantial infrastructure release may have a narrowly demonstrable impact on the GxP function.

Class	Definition and examples	Assessment and testing	Approval, monitoring and rollback
0: Non-impacting	A change shown not to affect the GxP function, its interpretation or its controlled configuration; for example, an editorial correction to non-operational documentation.	Record the change and the basis for no GxP impact. No new validation test is normally needed.	System or process owner closes the assessment under the applicable procedure; QA involvement follows local governance. Routine monitoring continues.
1: Bounded low-impact	An effect is possible but confined to known requirements or interfaces; for example, a narrowly scoped display change or technical patch with a stable interface.	Document GxP impact and risk; test affected functions and credible adjacent effects. Confirm the scope remains bounded.	Process owner and appropriate technical SME approve; QA reviews according to criticality. Confirm operation after release; provide a practical back-out route where needed.
2: Material performance-affecting	Behaviour of a GxP-relevant AI function may change, while intended use and the wider assurance baseline remain applicable; examples include fine-tuning, changed embeddings, material prompts or an updated model version.	Update the risk assessment and affected requirements; perform benchmark-based regression, new-failure-mode and relevant end-to-end testing. Document which previous evidence is reused.	QA, process owner and domain and technical SMEs approve. Use controlled release, enhanced monitoring and predefined rollback criteria. Obtain supplier evidence as appropriate.
3: Fundamental	Intended use, autonomy, important risk controls or the basis of previous evidence changes; examples include a new population, removal of human review or an unbounded model-family replacement in a critical workflow.	Establish a new or substantially revised validation baseline. Reassess supplier, architecture, requirements, risks and critical functions comprehensively.	Senior accountable owners and QA approve under company governance; assess any separate regulatory implications. Stage deployment and test fallback or rollback.

A model-family replacement is not automatically Class 3. It can be Class 2 if its effects can be bounded, the intended use and control architecture remain stable, and a credible bridge to the existing validation is possible. Conversely, changing a prompt may warrant Class 3 if that prompt changes a critical decision role or invalidates the previous evidence. The class must follow risk and uncertainty, not component names.

The classification should also be revisited when evidence changes. A proposed Class 1 change that produces unexpected critical errors in testing is no longer a low-impact change simply because it entered the process with that label.

How Much Revalidation Is Necessary?

The classification informs, but does not mechanically dictate, one of five responses.

No additional validation activity is appropriate when documented triage establishes that a change has no effect on the validated GxP function or configuration. It does not mean no documentation. The record must show what was changed, why it has no relevant impact and who accepted that conclusion.

Documented assessment only may be sufficient for a technically real but demonstrably non-functional change. The assessment should identify affected components, requirements and risks and explain why testing would not add meaningful assurance. This response should not become a default label for changes that are merely difficult to test.

Limited regression testing addresses a bounded change with known effects. Test the affected requirement, its interfaces and credible adjacent failure paths. The file should still show why broader evidence remains valid.

Partial revalidation is the likely response to a material AI performance change within a stable intended use. It may include a revised risk assessment, representative benchmark testing, critical-error and subgroup analyses, interface checks, evaluation of human review and a validation report that bridges unaffected evidence. Partial does not mean informal: for the affected function, its evidence may need to be rigorous.

Full revalidation is appropriate when a coherent bridge to the existing baseline cannot be made. Triggers include materially changed intended use or population; increased autonomy; removed or weakened human review; fundamental architecture change; an unbounded model replacement; widespread changes to critical requirements; inability to identify the effective production configuration; or failure of predefined criteria followed by substantial redesign.

There is no fixed test count separating these responses. A defensible rationale identifies which earlier claims remain true, which are uncertain, and what evidence resolves the uncertainty. This follows the broader risk-proportionate direction of Annex 11, ICH quality risk management and relevant clinical-trial computerised-system guidance, rather than an invented AI-specific regulatory threshold. [1–3,6]

Model Changed: What Now?

[AUTHOR’S PROPOSED APPROACH] The following sequence is designed for an SOP, change-record template or figure. Each step should leave an auditable decision.

1. Identify the change and production configuration. Establish what changed, when, where and for which transactions. Obtain the effective model, endpoint, prompt, retrieval and relevant software versions. If a critical production configuration cannot be identified, contain the change and investigate before relying on affected outputs.

2. Confirm the validated boundary. Determine whether the altered element can affect a validated GxP requirement, data flow, record, risk control or user action. A change outside that boundary may close after documented triage; an uncertain boundary calls for investigation, not an assumption of no impact.

3. Reconfirm intended use and context of use. Ask whether users, populations, source data, autonomy, downstream decisions or consequences have changed. A material change points towards a new validation baseline.

4. Assess potential harm and detectability. Consider product quality, patient or participant safety, pharmacovigilance obligations, data integrity and decision reliability. Assess whether qualified human review is likely to intercept an error before it matters.

5. Decide whether impact is bounded. Map affected requirements, interfaces and failure modes. If the effects cannot be bounded, expand the investigation and validation scope.

6. Approve an evidence plan before execution. Specify the reference data, test cases, metrics, acceptance limits, statistical treatment, human-workflow assessment and supplier evidence needed. Define the basis for reusing prior validation.

7. Evaluate the candidate configuration. Compare results with the approved baseline, investigate deviations and newly introduced errors, and determine whether every critical acceptance condition is met. Failure means rejection, remediation, restriction or a revised validation plan, not silent acceptance.

8. Approve controlled release. The accountable process owner accepts operational suitability and residual risk; QA and relevant SMEs approve according to the change class and quality system. Confirm that the tested and deployed configurations match.

9. Monitor and retain a fallback. Apply enhanced monitoring proportionate to risk and define conditions for rollback, suspension or further investigation. Reassess the cumulative effect of changes at periodic review.

This sequence also provides an answer to the apparent paradox of a stable UI with a changed model. The UI is one component of the configuration, not proof of unchanged system performance.

Worked Example: PV Information Extraction

A supplier proposes replacing Version A of the LLM underlying a validated pharmacovigilance extraction application with Version B. The application assists case processors by extracting patient, reporter, suspect product, adverse-event, date and seriousness information from source documents. Qualified personnel review the results before final case processing. The intended use and visible workflow are proposed to remain unchanged.

Open the change before deployment. The supplier notice should identify both versions, the effective endpoint, planned deployment date, reason for replacement, known changes to response format or safety behaviour, and availability of pre-production testing and rollback. The pharmaceutical company should compare architecture and configuration records rather than rely on a statement that the “API is compatible.” If the replacement has already occurred without notice, the issue becomes both a change and a deviation: establish the first affected date, identify affected records and assess whether retrospective review is needed.

Determine what is at risk. The PV process owner and SMEs should identify fields whose omission or misstatement could affect case recognition, seriousness, quality or timely processing. Risk is assessed at field and case level, considering the possibility that a reviewer misses a plausible-looking error. The fact that all outputs receive human review does not remove the need to establish that the review remains effective. EMA identifies PV as a potential AI application and expects marketing authorisation holders to validate, monitor and document model performance. [4]

Write a bridging plan. The company may reasonably retain earlier evidence for unaffected access controls, audit trails, infrastructure or standard operating procedures, provided it explains why each remains applicable. It should renew evidence for the changed extraction function, the response parser and interfaces, and any reviewer interaction that may be affected by different outputs. This is partial revalidation only if the prior assurance case remains coherent.

Govern the reference dataset. Freeze a representative benchmark with qualified, adjudicated labels and recorded provenance. It should include routine and difficult cases: short and long documents; relevant languages; multiple patients or events; follow-up information; negation; uncertain dates; poor-quality source material within the approved use; and serious cases. Preserve a locked core set for paired A-versus-B comparison and add fresh challenge cases directed at Version B’s plausible new failure modes. The formal set should not be used to iteratively tune B and then reported as independent confirmation. EMA stresses the importance of test-data independence in AI evaluation. [4]

Approve acceptance criteria in advance. Define absolute minimum performance on critical fields, limits on critical false negatives, acceptable structural or schema failures, and any non-inferiority margin relative to A. Include case-level completeness, relevant language or source-format subgroups, unsupported or fabricated information, and operational performance. The thresholds must follow the intended use and potential harm; there is no universal regulatory number to insert into this protocol. A single overall F1 score cannot compensate for a consequential increase in missed serious cases.

Run paired and repeated tests. Execute A and B against the locked core under matched, recorded settings. Retain source, raw response, parsed result and adjudicated outcome. Analyse discordant cases: did B correct an A error, preserve it or create a new one? Compare performance with the reference labels as well as with A; agreement between models is not proof that either is correct. For a stratified subset, repeat runs to measure variability in critical fields and unsupported content. Present uncertainty, including confidence intervals where

suitable, rather than treating a point estimate as exact.

Test what the old protocol may have missed. Challenge B for fabricated patient or event details, confusion between multiple cases, incorrect negation or chronology, omitted seriousness information, long-document truncation, inappropriate refusal, prompt injection contained in source material and changes to the output schema. Test error handling when the service is unavailable or returns incomplete data.

Evaluate the real reviewer control. Where feasible, have qualified reviewers assess representative A and B outputs without knowing which version produced them. Measure critical corrections, missed model errors, time required and unexpected changes to the review process. Investigate a more fluent answer that reviewers accept too readily as carefully as an obviously malformed answer.

Conclude and release. The validation report should identify the tested configuration, protocol and deviations; explain reference-data governance; present paired, subgroup and critical-error results; document residual risks; and state why reused evidence remains applicable. PV leadership, the process owner, technical SMEs and QA should approve according to the company's governance. At deployment, reconcile the production version with the tested version, retain the prior configuration or alternative fallback where feasible, and communicate any changed review instructions.

Monitor more closely after release. Track critical-field corrections, missed minimum case information, seriousness corrections, fabricated details, invalid structured responses, reviewer overrides and shifts in incoming document types. Set review intervals and alert thresholds in advance. An unidentified version change, material critical-error increase, loss of traceability or breach of an approved performance boundary should trigger investigation and potentially suspension or rollback.

This package need not repeat every test from the application's original validation. Its strength lies in a clear bridge: unaffected controls retain applicable evidence; affected functions receive new, risk-focused evidence; and monitored operation checks whether the pre-release conclusion remains true.

The Vendor-Control Problem

Externally hosted AI separates operational responsibility from technical control. The pharmaceutical company remains accountable for its GxP use, yet a foundation-model provider may control the model, endpoint routing, safety layer, subcontractors and retirement schedule. The contract and technical architecture must close as much of that gap as the intended use requires. Current Annex 11 already addresses supplier agreements and oversight; the revised Annex 11 consultation draft discusses more explicit arrangements for release of new versions and an opportunity for the regulated user to test before release. The latter remains a draft provision. [1,9]

For a GxP-relevant service, the supplier arrangement should cover the following minimum control questions:

Control question	Contractual or technical expectation
What changed?	Advance notification of material model, prompt, guardrail, retrieval, API and infrastructure changes; release notes and emergency-change notification.
What ran?	Stable, unique identification of the effective model or deployment and relevant endpoint version for each regulated transaction.
What can be tested?	Pre-production access, reasonable assessment time, supplier performance evidence, known limitations and support for customer testing.
Can the release be controlled?	Version pinning where available, agreed migration windows, previous-version availability or an alternative fallback, and rollback criteria.
What happens after a failure?	Prompt incident and degradation notification; investigation support; access to relevant records and remediation evidence.
Who else is involved?	Disclosure and control of material subprocessors and subcontractors, including changes affecting data handling or service delivery.

Control question	Contractual or technical expectation
Are records and data protected?	Defined processing locations, permitted data use, retention, access controls, cybersecurity responsibilities and breach response.
Can the company maintain oversight?	Risk-appropriate audit rights or credible alternative assurance, inspection support and access to relevant quality information.
What if the service ends?	Business-continuity arrangements, export of records and configuration, version-retirement notice and an exit strategy.

“Silent” model updates are particularly difficult because the company may not know when its tested configuration stopped operating. Behavioural monitoring can help detect unexpected change, but it may identify the problem only after regulated outputs have been produced. Monitoring is a useful independent control, not a complete substitute for change notification and version information.

[AUTHOR’S PROPOSED MINIMUM] For GxP-critical use, the retained transaction record should permit an investigator to associate an output with the effective model or deployment, endpoint, controlled instructions, relevant retrieval configuration, time, input, raw response, downstream transformation and human disposition. The precise metadata may differ by architecture. The requirement is functional: reconstruct what produced the regulated result and how that result was accepted or corrected.

A supplier need not disclose proprietary model weights for every use to be viable. Sufficient black-box testing, change information and operational controls may support an assurance case. But if the supplier cannot identify the effective version, provide a meaningful opportunity to assess material changes or support investigation of affected outputs, the company should not assume that an initially successful validation remains valid indefinitely. For a critical use, it may need to restrict the system’s role, add independently demonstrable controls, change supplier or suspend use.

Inspection Readiness

The inspection file for a Model A to Model B replacement should answer one question plainly: Why did the company conclude that the released Version B system remained fit for its approved GxP use without repeating the entire initial validation?

Keep the following evidence accessible and mutually consistent:

- An approved change request, supplier notification and exact identification of A, B, endpoints and effective dates.
- The approved intended-use and context-of-use statements, with confirmation of what remained unchanged.
- An architecture and configuration comparison, including prompts, retrieval, interfaces and human workflow.
- The GxP impact and risk assessments, change class, affected requirements and newly considered failure modes.
- A protocol approved before formal testing, with the validation scope and reasons for reusing specific prior evidence.
- Benchmark provenance, version, reference-label method, independence controls and representative-case rationale.
- Predefined acceptance criteria, paired A-versus-B findings, critical-error review, relevant variability analysis and investigated deviations.
- Evidence that human review and downstream interfaces remain effective.
- The validation report and approvals from QA, the process owner and required SMEs.
- Proof that the deployed configuration matches the tested configuration.
- Release, fallback or rollback records, enhanced-monitoring criteria and subsequent results.

A clean test summary without a configuration record is incomplete. So is a signed change assessment that cannot explain why the benchmark represented the actual GxP workload. Inspection readiness depends on the continuity of the argument from intended use, through risk and evidence, to the configuration that operated in production.

Where the Boundary Remains Unclear

The regulatory direction is clear enough to require control and justification, but several practical thresholds remain unsettled. No harmonised cross-GxP rule defines exactly when an AI model update moves from focused testing to partial or full revalidation. There is likewise no universal acceptable hallucination rate, false-negative limit or statistical non-inferiority margin. Each must be defined in relation to the task and the consequences of failure.

Black-box suppliers create a second uncertainty. Performance testing may establish credible fitness without disclosing weights or complete training data, yet supplier secrecy cannot excuse the absence of an identifiable production configuration or a means of investigating errors. The appropriate minimum disclosure will differ between a fully reviewed drafting aid and an application that autonomously filters safety-relevant information. EMA’s reflection paper accommodates risk-based evidence rather than demanding a single model type, while still placing responsibility for fitness and monitoring on the regulated organisation. [4]

Continuous-learning models pose a distinct challenge because the production configuration can change as it is used. Proposed Annex 22 takes a restrictive position for its specified critical GMP manufacturing scope; that draft position should not be extrapolated into a settled rule for every pharmaceutical domain. Where adaptation is contemplated, change boundaries, approval points, data governance and the ability to identify the effective state become particularly important. [10]

RAG and agentic systems further complicate configuration control: retrieved content, indexes, tools, permissions and external services can change independently of the foundation model. Human oversight is also hard to quantify over time; reviewer familiarity, workload and trust can alter its effectiveness. Finally, exact reproduction of a historical hosted-model response may be impossible after a version is retired. That makes preservation of contemporaneous inputs, outputs, configuration metadata and review decisions more important, not less.

The dividing line between an “updated system” and a “new system” is therefore not a product-name or UI decision. It is an evidence decision: can the company still connect its approved intended use and risk controls to the configuration now operating, using a credible bridge from the earlier validation?

Conclusion

Pharmaceutical organisations do not need to choose between freezing AI technology indefinitely and completely revalidating an entire system after every update. They do need a disciplined way to distinguish a bounded change from one that defeats the previous assurance case.

[AUTHOR’S PROPOSED APPROACH] Each GxP AI application should have a Predetermined AI Change Control Plan: a prospective description of foreseeable changes, the information the supplier must provide, factors that determine change class, evidence expected at each level, acceptance boundaries, approval roles, monitoring and fallback. The term borrows a useful lifecycle-control idea from medical-device PCCPs. It is proposed pharmaceutical good practice, not an existing general legal requirement for pharmaceutical GxP systems. [15,16]

The durable sequence is:

Initial Validation → Controlled Configuration → Change Detection → Impact Assessment → Proportionate Revalidation → Approved Release → Continuous Performance Monitoring

At each transition, the company should be able to show what it knew, what it tested, what it accepted and what it would do if performance deteriorated. That is how the engine can change while the validated state remains defensible.

References

1. European Commission. EudraLex Volume 4, Annex 11: Computerised Systems. Revision 1; operative from 30 June 2011.

2. International Council for Harmonisation. ICH Q9(R1): Quality Risk Management. Final guideline.

3. International Council for Harmonisation. ICH Q10: Pharmaceutical Quality System. Final guideline.

4. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. Final reflection paper; 2024.

5. International Council for Harmonisation. ICH E6(R3): Guideline for Good Clinical Practice. Final Principles and Annex 1.

6. European Medicines Agency. Guideline on computerised systems and electronic data in clinical trials.

7. Organisation for Economic Co-operation and Development. Application of GLP Principles to Computerised Systems, Advisory Document No. 17.

8. Council for International Organizations of Medical Sciences. Artificial Intelligence in Pharmacovigilance: Report of CIOMS Working Group XIV. Published December 2025.

9. European Commission. Stakeholder consultation on revised EU GMP Chapter 4 and Annex 11 and proposed Annex 22. Opened 7 July 2025; closed 7 October 2025. Consultation material, not operative revisions.

10. Pharmaceutical Inspection Co-operation Scheme. Proposed Annex 22: Artificial Intelligence. Consultation draft, 2025.

11. US Food and Drug Administration. Computer Software Assurance for Production and Quality Management System Software. Final guidance; issued 3 February 2026; medical-device production and quality-management-system scope.

12. US Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products. Draft guidance; January 2025.

13. International Society for Pharmaceutical Engineering. GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems, Second Edition. Industry guidance.

14. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Voluntary framework; 2023.

15. European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. EU AI Act; application must be assessed according to its scope and current amendments.

16. US Food and Drug Administration. Predetermined change control plans for AI-enabled device software functions: final-guidance information. Medical-device context; not a general pharmaceutical GxP requirement.

About the author

Reinhold Schilling writes on the practical governance of AI in pharmaceutical GxP systems, with an emphasis on validation, pharmacovigilance, supplier oversight and inspection-ready change control.