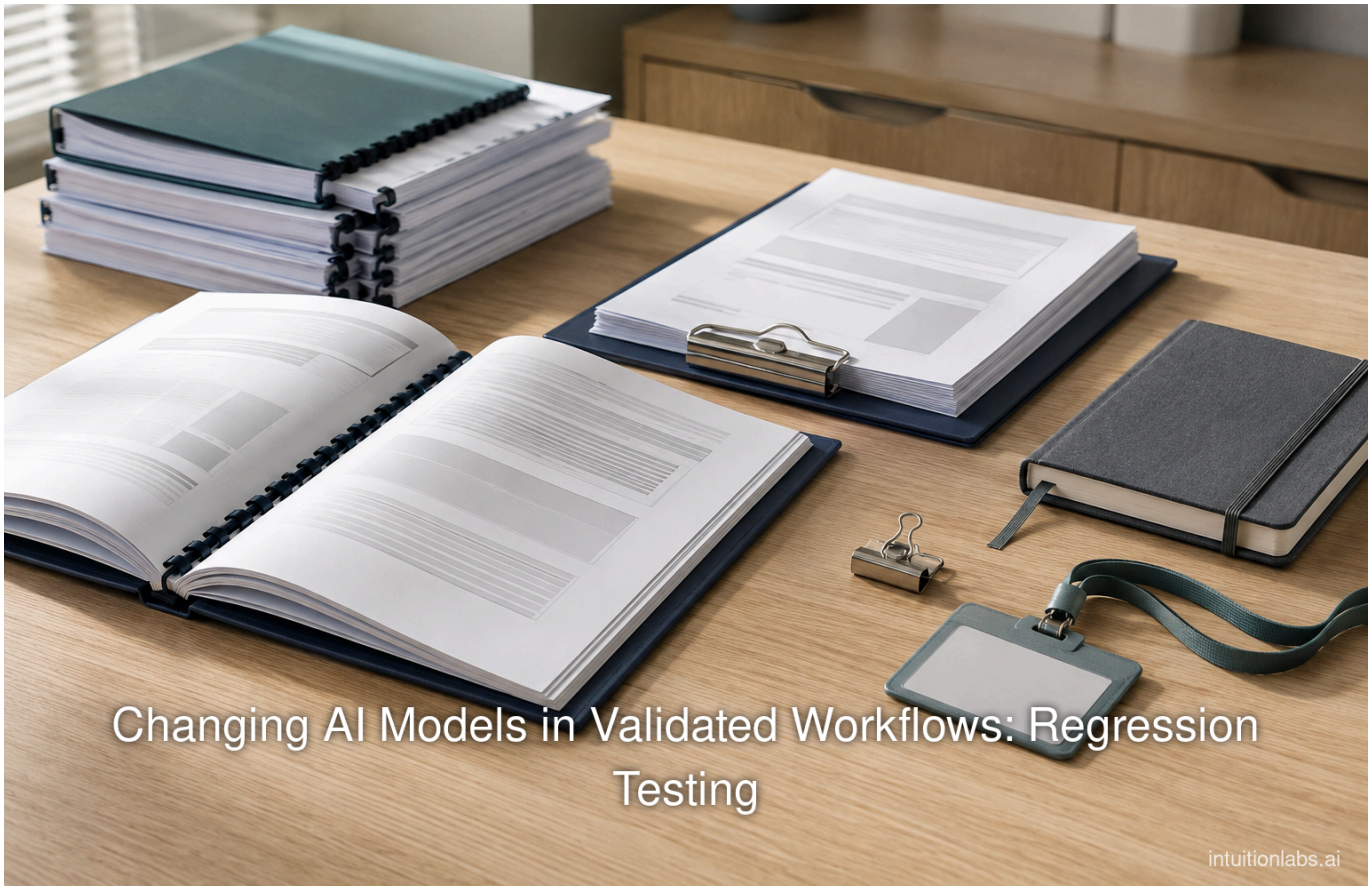
Changing AI Models in Validated Workflows: Regression Testing

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · regression testing · ai model change control · validated workflows · gxp compliance · computer system validation · gamp 5 · llm versioning · change management · fda csa guidance · model deprecation policy



Executive Summary

When an AI model underneath a validated computerized system changes, whether through a routine vendor version bump, a deprecation notice, or a deliberate provider switch, no GxP (Good x Practice) regulation currently uses the words "AI model" or "large language model (LLM)" in its binding change-control text. Organizations instead extend decades-old rules, the United States' 21 CFR Part 11 and the European Union's GMP Annex 11, whose requirements have distinct scopes (www.ecfr.gov) (health.ec.europa.eu), to a genuinely new kind of change: a probabilistic model whose behavior can shift between dated versions in ways a deterministic software patch never would. The clearest regulatory template is the U.S. Food and Drug Administration's (FDA) Predetermined Change Control Plan (PCCP) framework for AI-enabled medical devices, finalized in 2024 and requiring a pre-specified Description of Modifications, Modification Protocol with acceptance criteria, and Impact Assessment before a change can bypass a new marketing submission (www.fda.gov), but that framework applies narrowly to authorized devices, not the broader universe of validated pharmaceutical workflows that merely use an LLM as a component.

The practical mechanism forcing a decision is vendor versioning policy. Anthropic commits to at least 60 days' notice before retiring a publicly released model (platform.claude.com); OpenAI commits to at least six months for generally available models but as little as two weeks for preview releases (platform.openai.com) (platform.openai.com); and Microsoft's Azure AI Foundry applies a 60-day notice and a shortened 12-month lifecycle to third-party hosted models (learn.microsoft.com) (learn.microsoft.com). Version-specific regression testing remains a prudent engineering practice before a production cutover. FDA's Computer Software Assurance guidance is nonbinding and applies to software used as part of medical-device production or the quality management system (www.fda.gov).

This report separates regulatory obligation from recommended practice: Part 11 applies to records in electronic form that are created, modified, maintained, archived, retrieved, or transmitted under agency record requirements, while Annex 11 applies to computerised systems used as part of GMP-regulated activities. Documented change control and risk-scaled effort may be requirements only within an applicable framework and scope; dated-version pinning, automated CI regression suites, and shadow or blue-green rollout before cutover remain recognized but non-mandated engineering practices (platform.claude.com) (docs.langchain.com) (docs.aws.amazon.com). It closes with a worked change-control record and a bounded conclusion: the compliance gap will likely narrow as bodies like ISPE publish AI-specific guidance, but until then, risk-based, documented, test-before-you-switch discipline is the most defensible available approach.

Introduction and Background

Enterprise deployments of large language models (LLMs) rarely stay on one model version for long. Providers retire, rename, and replace models on cycles measured in months, and every one of those changes can alter the output of a workflow that a life sciences organization has already validated for a regulated purpose. This report addresses a narrow but increasingly common operational question: when the AI model underneath a validated computerized system changes, what regression testing, documentation, and change control are actually required, and which practices are merely good engineering hygiene rather than a regulatory mandate. The distinction matters because treating every model update as a full revalidation event is unaffordable at the cadence vendors now ship updates, while treating no update as consequential risks silently degrading a system that a health authority, an auditor, or a patient depends on.

The regulatory backdrop is not new. Electronic-record and computerized-system rules such as the United States' 21 CFR Part 11 and the European Union's Good Manufacturing Practice (GMP) Annex 11 address computerized systems within their respective scopes, well before generative AI existed (www.ecfr.gov) (health.ec.europa.eu). What is new is the object of change: a **large language model (LLM)**, an AI system trained to generate text, code, or structured output from natural-language prompts, whose behavior can shift between versions in ways that are harder to predict than a traditional software patch. The U.S. Food and Drug Administration (FDA) has separately built an entire regulatory vocabulary, most visibly the **Predetermined Change Control Plan (PCCP)** framework for AI-enabled medical devices, specifically to handle this class of problem (www.fda.gov).

This report treats the question in three parts: the mechanics that force a regression-testing decision (how providers version, pin, and deprecate models); what GxP (Good x Practice, the umbrella term for GMP, Good Clinical Practice, and related pharmaceutical quality regulations) rules and FDA AI/ML guidance actually require versus what is industry-recommended engineering practice with no regulatory citation behind it; and a realistic change-

control document flow from triggering event to rollback plan. All version identifiers and policy figures below are anchored to their observation date because vendor deprecation policies change quickly; readers should re-confirm current policy pages before acting, as of September 2026.

Key Changes

Version Pinning and Model Identity

The first practical control is knowing precisely which model produced a given output, and being able to keep using that exact model on purpose. Major providers now expose this as a **dated, pinned model identifier**: a version string that never silently changes its underlying weights once published. Anthropic states that for its Claude models, "the underlying model remains constant for the lifetime of that ID" (platform.claude.com), and structures dated snapshots using a `claude-{name}-{major}-{minor}-{YYYYMMDD}` naming pattern (platform.claude.com). Critically, Anthropic also warns that a dateless alias is not automatically safe to treat as pinned: "dateless model IDs such as" the generation-4.6-plus aliases "behave as evergreen pointers" that can silently move to a newer underlying model over time (platform.claude.com), and that even a genuinely pinned model ID is not a perfect guarantee of identical behavior, because "infrastructure updates produce minor differences in observable behavior even when the model ID and weights have not changed" (platform.claude.com).

The other three major API providers use structurally similar but distinctly-worded schemes, summarized fully in *Table 1* below. OpenAI distinguishes "generally available" (GA) models, which receive at least six months' deprecation notice, from preview or specialized variants that "may be retired with much shorter notice, such as 2 weeks" (platform.openai.com) (platform.openai.com), and formally separates "deprecated" (announced, still working) from "legacy" (still working but receiving no further updates) states (platform.openai.com). Microsoft's Azure AI Foundry (formerly Azure OpenAI Service) commits to at least 60 days' notice for GA-model retirement, and, notably, applies a shortened 12-month total lifecycle (versus the standard 18 months) to third-party models it hosts from Anthropic, DeepSeek, Fireworks, and Mistral AI (learn.microsoft.com) (learn.microsoft.com). Google's Gemini API defines three alias tiers, Stable, Preview, and Latest, recommending that "most production apps should use a specific stable model" rather than a floating alias (ai.google.dev) (ai.google.dev), while Google Cloud's Vertex AI platform commits that published retirement dates "won't be moved to an earlier date than what is listed" (cloud.google.com) (cloud.google.com).

Table 1 below summarizes these four providers' publicly documented model lifecycle and notice policies as of September 2026.

Provider	Minimum notice, GA models	Preview/short-lifecycle notice	Auto-upgrade behavior
Anthropic (Claude)	At least 60 days for publicly released models, with email plus documentation notice (platform.claude.com) (platform.claude.com)	Dateless aliases can move to a new model without a discrete retirement event (platform.claude.com)	Dated snapshot IDs never change weights; dateless aliases behave as evergreen pointers (platform.claude.com) (platform.claude.com)
OpenAI	At least 6 months for GA models (platform.openai.com)	As little as 2 weeks for preview/specialized variants (platform.openai.com)	No silent upgrade of a pinned dated snapshot; enterprise customers may

Provider	Minimum notice, GA models	Preview/short-lifecycle notice	Auto-upgrade behavior
			provision dedicated capacity past a shutdown date (platform.openai.com)
Microsoft Azure AI Foundry	At least 60 days GA, at least 30 days preview (learn.microsoft.com)	Third-party models (Anthropic, DeepSeek, Fireworks, Mistral AI) follow a 12-month total lifecycle versus 18 months standard (learn.microsoft.com)	Standard/Global Standard deployments auto-upgrade at retirement unless set to `NoAutoUpgrade`; Provisioned deployments never auto-upgrade (learn.microsoft.com)
Google (Gemini API / Vertex AI)	At least 2 weeks' email notice before a "latest" alias changes (ai.google.dev); most models available 12+ months (cloud.google.com)	Short-term-availability models retire 45 days after a replacement ships (cloud.google.com)	Stable aliases are recommended for production; retirement dates may extend but never move earlier (ai.google.dev) (cloud.google.com)

The pattern across all four vendors is the same underlying design: a dated or otherwise pinned identifier is the unit a validated workflow should reference in its own configuration and change-control records, not a bare model family name. A validated system that calls a floating alias has, in effect, delegated its own change control to the vendor's release cadence, which is a defensible engineering choice only if that delegation is itself documented and risk-assessed, not a default.

Change Impact Assessment

For computerised systems used in GMP-regulated activities, Annex 11 calls for risk management throughout the system lifecycle and controlled changes under a defined procedure. EU GMP Annex 11 states the governing principle plainly: introducing or changing a computerised system must ensure "there should be no resultant decrease in product quality, process control or quality assurance" (health.ec.europa.eu), and ties "the extent of validation and data integrity controls" to "a justified and documented risk assessment" (health.ec.europa.eu). Annex 11's change-control clause requires that "any changes to a computerised system including system configurations should only be made in a controlled manner in accordance with a defined procedure" (health.ec.europa.eu), and separately requires that validated systems be "periodically evaluated to confirm that they remain in a valid state" (health.ec.europa.eu). In the United States, 21 CFR Part 11 requires "validation of systems to ensure accuracy, reliability, consistent intended performance" and includes revision and change-control procedures for an audit trail documenting development and modification of systems documentation (www.ecfr.gov) (www.ecfr.gov).

Neither Part 11 nor Annex 11 names AI or LLMs specifically; both apply to model changes only by extension of their general change-control language. ICH Q9(R1), the International Council for Harmonisation's Quality Risk Management guideline (adopted 18 January 2023), supplies the scaling principle organizations use to fill that gap: "the level of effort, formality and documentation of the quality risk management process should be commensurate with the level of risk" (database.ich.org) (database.ich.org). Applied to a model swap, a low-risk internal drafting assistant and a high-risk model embedded in a batch-release decision warrant categorically different depths of impact assessment.

FDA's clearest, most model-specific impact-assessment framework exists in its medical device rules for continuously updating algorithms. Its 2019 discussion paper on AI/ML-based Software as a Medical Device (SaMD) formally defines a "locked" algorithm as one that "provides the same result each time the same input is applied to it

and does not change with use," contrasted with an adaptive algorithm whose "output may be different before and after the changes are implemented" (www.fda.gov) (www.fda.gov). The same paper states SaMD "exist on a spectrum from locked to continuously learning" (www.fda.gov), and that historically, "algorithm changes likely require FDA premarket review for changes beyond the original market authorization" (www.fda.gov). FDA's subsequent PCCP guidance operationalizes a path around that constraint: a PCCP must contain a "Description of Modifications," a "Modification Protocol," and an "Impact Assessment" (www.fda.gov), where the Description of Modifications specifies "the characteristics and performance of the planned modifications" (www.fda.gov) and the Modification Protocol defines "verification and validation testing activities that will support the planned modifications and acceptance criteria" (www.fda.gov). An authorized PCCP lets a manufacturer implement a pre-specified change "without triggering the need for a new marketing submission" (www.fda.gov); deviating from it "would generally cause the device to be adulterated and misbranded" under the Federal Food, Drug, and Cosmetic Act (www.fda.gov).

Table 2 below maps the frameworks discussed in this section to what each one actually requires with respect to a model or system change, and whether AI or LLM changes are named explicitly.

Framework	Governing body	Core change-control requirement	Names AI/LLM changes explicitly?
21 CFR Part 11	FDA	Part 11 addresses revision and change control for systems documentation within its electronic-record scope	No, general electronic-records rule
EU GMP Annex 11	European Commission	Risk-assessed, controlled-procedure change management; periodic re-evaluation of validated state	No, general computerised-systems annex
ICH Q9(R1)	ICH	Risk-management effort must scale to the level of risk	No, general quality risk management
FDA Computer Software Assurance guidance	FDA (CDRH/CBER)	Risk-based, "least-burdensome" assurance strategy tailored to each software feature's intended use (www.fda.gov)	No, but explicitly covers production/QMS software changes
FDA PCCP guidance	FDA (device centers)	Pre-authorized Description of Modifications, Modification Protocol with acceptance criteria, and Impact Assessment	Yes, written specifically for AI-enabled device software
ISPE GAMP Guide: Artificial Intelligence	ISPE (industry body, not a regulator)	Risk-based lifecycle guidance for AI-enabled GxP computerized systems (ispe.org)	Yes, dedicated AI guide (paid publication)

The practical reading of this table is that no GxP regulation currently uses the words "AI model" or "LLM" in its binding change-control text; organizations are extending decades-old computerized-system change-control language to a new kind of change. FDA's device-specific PCCP and GMLP frameworks are the closest thing to an explicit regulatory answer, but they apply to authorized medical devices, not to the broader universe of validated pharmaceutical, clinical, and quality workflows that merely use an LLM as a component.

Acceptance Criteria and Regression Test Design

An impact assessment identifies that testing is needed; acceptance criteria and a test design determine whether the new model version actually passes. As described above, FDA's PCCP guidance requires that a Modification Protocol specify, in advance, both the testing activities and the acceptance criteria a change must meet, which is the regulatory articulation of a general principle: acceptance criteria must exist before the new model is run, not be reverse-engineered from whatever the new model happens to produce.

Acceptance criteria should be built around measured, task-specific outputs rather than a vendor's general capability claims. **Item-level regression testing** compares specific input-output pairs across versions rather than relying on a single aggregate pass rate.

Model providers and evaluation-tooling vendors converge on the same engineering pattern for building that comparison: a fixed test set ("golden dataset") run against both the old and new model, with automated pass/fail scoring. OpenAI's own open-source Evals framework states its purpose is to "understand how different model versions might affect your use case" (raw.githubusercontent.com). LangChain's LangSmith platform defines "regression testing" as a distinct evaluation use case meant to "ensure new versions don't degrade quality" (docs.langchain.com), and recommends versioning the test dataset itself so a pipeline can "target specific versions in CI pipelines to ensure dataset updates don't break workflows" (docs.langchain.com). Braintrust's evaluation documentation frames the same practice as a way to "detect regressions before they reach production" (www.braintrust.dev), recommending that teams "run evals on every pull request to catch regressions" (www.braintrust.dev). None of this tooling is GxP-specific or mandated by any regulation; it is industry engineering practice that happens to produce exactly the evidence a GxP acceptance-criteria record needs.

Effective acceptance criteria for a model swap in a validated workflow typically combine:

- **A fixed, representative test set** covering the workflow's actual production input distribution, not a generic public benchmark.
- **Item-level comparison** between old-model and new-model outputs on that test set, not only an aggregate score.
- **Task-specific quantitative thresholds** (accuracy, factual-consistency rate, formatting compliance) defined before testing begins, mirroring the pre-specification principle described above.
- **Qualitative human review** of a sampled subset, particularly for open-ended generation tasks where automated scoring is unreliable.
- **Documented pass/fail disposition** tied to the specific dated model identifier being evaluated, not the model family name.

Audit Evidence and Documentation

For Annex 11 computerised systems used as part of GMP-regulated activities, validation documentation includes change-control records if applicable. Annex 11's change-control clause requires changes to be made in accordance with a defined procedure, and Part 11 separately requires "revision and change control procedures to maintain an audit trail that documents time-sequenced development and modification of systems documentation" (www.ecfr.gov). FDA's Good Machine Learning Practice (GMLP) principles, published jointly with Health Canada and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) in October 2021, extend this expectation specifically to machine learning systems: Principle 10 states that "when models are periodically or

continually trained after deployment, there are appropriate controls in place to manage risks of overfitting, unintended bias, or degradation of the model (for example, dataset drift)" (www.fda.gov), and Principle 1 requires that "multi-disciplinary expertise is leveraged throughout the total product life cycle," not only at initial development (www.fda.gov).

For a model swap specifically, the minimum audit-evidence package should include: the triggering event (vendor deprecation notice, planned upgrade, or provider switch) with its date; the impact assessment and risk classification; the pre-specified acceptance criteria; the regression test results, item-level where feasible; the approval record, including who reviewed and authorized the change; and the effective date the new model version went live. NIST's AI Risk Management Framework (AI RMF 1.0, published January 2023) reinforces the same point from a general AI-governance angle, stating that "post-deployment AI system monitoring plans are implemented" as part of ongoing risk management, including "change management" (nvlpubs.nist.gov), and noting that AI systems "require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift" than traditional software (nvlpubs.nist.gov).

Rollback and Contingency Planning

A regression test that fails, or a production issue discovered after a model swap has gone live, both require a pre-planned way back. Two documented technical patterns reduce this risk: shadow testing and staged (blue-green or canary) rollout. AWS SageMaker's shadow-testing feature lets a team "validate changes to any component of your production variant, namely the model, the container, or the instance, without any end user impact," by routing a copy of live traffic to the candidate version while the incumbent version keeps serving real users (docs.aws.amazon.com), recommended specifically when "promoting a new model that has been validated offline to production" but still needing to check "latency and error rate" (docs.aws.amazon.com). Microsoft's Azure Machine Learning documents an equivalent staged, "blue-green" pattern, instructing teams to "mirror a percentage of live traffic to the green deployment to validate it" before shifting production traffic over (learn.microsoft.com).

At the vendor level, rollback capacity is bounded by how long the old model version remains callable. Anthropic provides "at least 60 days' notice before model retirement for publicly released models" and states "impacted customers will always be notified by email and in the documentation" (platform.claude.com) (platform.claude.com), which functions as a rollback window. Microsoft's Azure AI Foundry states that "Provisioned deployments are NOT auto-upgraded" at retirement, meaning a deployment stops working unless migrated in advance, while Standard and Global Standard deployments auto-upgrade unless the customer opts out (learn.microsoft.com); Microsoft also states plainly that "retirement dates aren't extendable" (learn.microsoft.com). A rollback plan should therefore specify, in writing, before the swap: the exact prior dated model identifier, how long it remains available per vendor policy, the rollback trigger conditions, and who is authorized to execute it.

Implementation Considerations and Process Changes

Translating the controls above into an organization's standard operating procedures (SOPs) requires separating what a regulation actually obligates from what is prudent but optional engineering practice. Conflating the two produces two opposite failure modes: over-validating routine, low-risk model refreshes at a cost the organization cannot sustain, or under-validating a change that materially affects a regulated output because no named rule appeared to require more. *Table 3* below draws that line explicitly for the controls discussed in this report.

Practice	Regulatory requirement or recommended engineering practice	Basis
Documented change control for applicable computerized-system changes	**Regulatory requirement** (GxP)	Annex 11 clause 10; 21 CFR 11.10(k)(2) requires revision and change-control procedures for systems documentation (health.ec.europa.eu) (www.ecfr.gov)
Risk-based scaling of validation effort	**FDA CSA guidance: nonbinding; limited to medical-device production or quality-management-system software**	FDA CSA guidance (www.fda.gov)
Pre-specified acceptance criteria before a change is tested	**Regulatory requirement** for AI-enabled devices under an authorized PCCP; recommended practice elsewhere	FDA PCCP guidance
Post-deployment monitoring for model drift	**Nonbinding best-practice guidance** for AI/ML medical-device development (FDA/Health Canada/MHRA GMLP Principle 10); recommended elsewhere	FDA/Health Canada/MHRA GMLP
Pinning a dated model identifier instead of a floating alias	**Recommended engineering practice**, not mandated by any cited regulation	Vendor documentation only (platform.claude.com) (platform.claude.com)
Automated item-level regression test suites in CI/CD pipelines	**Recommended engineering practice**, not mandated by any cited regulation	LangSmith, Braintrust, OpenAI Evals documentation (docs.langchain.com) (www.braintrust.dev) (raw.githubusercontent.com)
Shadow testing or blue-green staged rollout before full cutover	**Recommended engineering practice**, not mandated by any cited regulation	AWS, Microsoft documentation (docs.aws.amazon.com) (learn.microsoft.com)

The clearest regulatory signal in favor of a risk-tiered process, rather than a single fixed procedure, comes from FDA's own Computer Software Assurance (CSA) guidance, finalized September 24, 2025 and reissued February 3, 2026 under the same docket as "Computer Software Assurance for Production and Quality Management System Software" (www.federalregister.gov) (www.fda.gov). The current, February 2026 version explicitly supersedes the September 2025 text (www.fda.gov), and defines computer software assurance as "a risk-based approach for establishing and maintaining confidence that software is fit for its intended use," applicable to "changes to software used in production or the quality management system" (www.fda.gov). Its central instruction is that "the burden of validation is no more than necessary to address the risk" (www.fda.gov), achieved by directing manufacturers to "examine the intended uses of the individual features, functions, and operations to facilitate development of a risk-based assurance strategy" rather than applying one uniform validation depth to every change (www.fda.gov), and it explicitly names "risk-based testing, unscripted testing, continuous performance monitoring, and data monitoring" as acceptable assurance methods (www.fda.gov). Within the guidance's medical-device production or quality-management-system scope, FDA recommends examining intended uses of individual software features, functions, and operations when developing a risk-based assurance strategy.

Change Control Record for a Model Swap: A Worked Example (Hypothetical Example)

The following illustrates, as a hypothetical example only, how the controls above compose into a single change-control record for a mid-risk validated workflow, such as a clinical-documentation drafting assistant, moving from one dated LLM version to another:

- **Trigger.** Vendor deprecation notice received, logged with the notice date and the vendor's stated retirement date.
- **Risk classification.** Quality assigns a risk tier based on the workflow's regulated function (here: mid-risk, output is human-reviewed before use).
- **Impact assessment.** Documents what changed, which system functions depend on it, and the risk rationale, mirroring the "Description of Modifications" and "Impact Assessment" components described above.
- **Acceptance criteria.** Defined before testing begins: minimum factual-consistency rate, maximum formatting-compliance failure rate, required human-reviewer sign-off rate, consistent with the pre-specification principle described above.
- **Regression test execution.** The fixed test set is run against both dated model identifiers; item-level differences are logged, not only the aggregate pass rate.
- **Shadow or canary check.** The new model runs against a slice of live production traffic in parallel with the incumbent model before full cutover (docs.aws.amazon.com) (learn.microsoft.com).
- **Approval and audit trail entry.** Quality and system-owner sign-off recorded with date, reviewer identity, and disposition (www.ecfr.gov).
- **Effective date and rollback plan.** New model identifier goes live on a recorded date; the prior identifier and its remaining vendor-availability window become the rollback path, with named trigger conditions and an accountable approver.

Data Analysis and Evidence

Quantifying how much a model version change can move outputs, and how the industry is responding to that instability, requires evidence from three different lines of research: academic measurement of version-to-version drift, evaluation-platform methodology, and survey data on enterprise AI governance and validation practice.

On the governance side, a Gartner survey of 360 respondents from organizations with at least 250 full-time employees, fielded May through June 2025, found that "organizations that perform regular audits and assessments of AI system performance and compliance are over three times more likely to achieve high GenAI value" than those that do not (www.gartner.com) (www.gartner.com), and separately that organizations "that provide GenAI ethics training are 1.7 times more likely" to report higher realized value (www.gartner.com). Kneat Solutions' "State of Validation" annual industry survey (2024 edition) found that "two-thirds of respondents" reported "validation expenses consume over eight percent of project budgets, often exceeding 10 percent," and that "balancing cost and resource constraints" was cited as "the main future challenge for most respondents (57 percent)" (3013089.fs1.hubspotusercontent-na1.net) (3013089.fs1.hubspotusercontent-na1.net); the same survey found "seventy percent of respondents believe AI and machine learning will play a pivotal role" in the future of validation practice (3013089.fs1.hubspotusercontent-na1.net), a forward-looking industry expectation rather than a measured outcome. Separately, FDA maintains a running, dated public list titled to "identify AI-enabled medical devices that are authorized for marketing in the United States," which the agency itself describes as "not a comprehensive

resource of AI-enabled medical devices" rather than an exhaustive count (www.fda.gov) (www.fda.gov); as of the page's own September 4, 2026 currency date, a manual count of the list's entries runs into the low thousands, though FDA does not publish that running total as a headline statistic itself, so it should be read as a lower bound on regulatory AI/ML device activity, not a precise market count.

Taken together, this evidence supports a specific, bounded claim: LLM outputs can move substantially between provider-issued versions, in ways that are not always visible in aggregate accuracy metrics, which is the direct empirical justification for item-level regression testing over version comparisons rather than trusting a single aggregate score, and it aligns with the general industry finding that more disciplined AI governance correlates with better realized value from generative AI investment.

Case Studies and Real-World Examples

A Documented Cross-Provider Model Migration

A published engineering account of a production migration between model providers, moving a workload from GPT-4.1 to Claude, describes the migration explicitly as "a real production-style experiment, not as a benchmark exercise" (blog.dhlabs.ai), and describes its pre-cutover validation step as "running both models on a subset of production traffic and comparing outputs side-by-side to catch any regression" (blog.dhlabs.ai), the same shadow-testing logic AWS and Microsoft formalize in their platform documentation. This is a single vendor engineering blog rather than a peer-reviewed or regulator-audited source, and it is presented here only as an illustration of the shadow-comparison methodology in practice, not as a benchmark of which provider performed better; readers should treat its specific performance claims as one team's self-reported experience rather than independently verified results.

Implications and Future Directions

Two structural pressures point toward more, not less, formalization of AI model change control in regulated environments over the next several years. First, GMLP Principle 10 presents post-deployment monitoring as nonbinding guidance for AI/ML medical-device development: deployed models have the capability to be monitored in real-world use, with controls for retraining risks (www.fda.gov). As more life sciences organizations embed LLMs into workflows adjacent to, though not necessarily inside, regulated medical devices, the PCCP-style discipline (pre-specified modifications, pre-specified acceptance criteria, and a documented impact assessment) is likely to be adopted informally as a best-practice template even where no device authorization requires it, simply because it is the most complete public template available.

Second, the vendor side of this problem is not standing still. Microsoft's decision to apply a shortened 12-month lifecycle specifically to third-party hosted models (learn.microsoft.com) suggests notice periods and lifecycle lengths are still being actively tuned by vendors, not settled industry norms. A validated workflow's rollback capacity is only as good as the shortest notice period among the vendors and model tiers it depends on, which argues for treating vendor deprecation-policy monitoring as a standing operational responsibility, not a one-time reading of a policy page at initial validation.

Consultancies operating in this space, including intuitionlabs.ai, describe their role as helping life sciences organizations operationalize governance around AI systems rather than supplying the underlying models themselves; the firm positions itself around "cutting-edge AI solutions designed specifically for pharmaceutical and life science organizations" (intuitionlabs.ai) and describes its flagship offering as helping clients "implement governed AI in one department, support users in real workflows" before scaling further (intuitionlabs.ai). That framing is consistent with the practical reality this report describes: the hard part of managing an AI model change in a validated workflow is the surrounding documentation, risk classification, and test design, which is process work an external quality or validation partner is positioned to support rather than a task a model vendor performs on a customer's behalf.

This report is narrowly scoped to the change-control mechanics of a model swap; broader questions of enterprise LLM deployment architecture, training rollout, and organizational adoption for a specific platform, such as Anthropic's Claude Enterprise, are addressed separately and are not restated here (intuitionlabs.ai). Looking forward, organizations should expect industry bodies to keep publishing more explicit guidance in this area. ISPE's dedicated "GAMP Guide: Artificial Intelligence," published July 2025, is one such signal, explicitly framed around "risk-based efforts to allow for efficient, compliant processes" for AI-enabled GxP systems (ispe.org), though as a paid, member-priced publication it is not yet as widely and freely accessible as the foundational GAMP 5 framework it extends.

Frequently Asked Questions (FAQs)

Does changing an AI model always trigger full computer system revalidation? Not necessarily. FDA's nonbinding CSA guidance recommends a risk-based approach for software used as part of medical-device production or the quality management system; applicable requirements for another workflow depend on its governing framework and scope.

What does "change control for AI model updates" mean under GxP? For a computerised system within EU GMP Annex 11's scope, changes including system configurations should be made in a controlled manner under a defined procedure. Whether and how that provision applies to a model version, provider, or configuration depends on the regulated activity and system scope.

How do you perform an AI model version change impact assessment in a life sciences setting? At minimum, document what changed (model identifier, provider, any known behavioral differences), which validated functions depend on it, a risk classification, and required testing, following the same Description-of-Modifications-and-Impact-Assessment structure described above.

What regression testing is appropriate when swapping the LLM behind a validated computerized system? A fixed, representative test set with pre-specified, quantitative acceptance criteria can support item-level comparison of old and new model outputs.

Is LLM model change management in pharma validation different from general software change management? The applicable change-control requirements should be assessed against the governing framework and system scope. Model pinning and regression testing are engineering practices that organizations may use to manage a model change.

What is the regulatory risk of switching LLM providers in a validated system? There is no dedicated regulation for "switching LLM providers" as such. For computerised systems used in GMP-regulated activities, Annex 11 provides that changes including system configurations should be made in a controlled manner under a defined procedure.

Does GAMP 5 address AI model change control specifically? The base GAMP 5 framework predates widespread LLM adoption and does not address AI-specific change control in its core text; ISPE has since published a dedicated "GAMP Guide: Artificial Intelligence" (July 2025) aimed at that gap, alongside earlier practitioner guidance in ISPE's own Pharmaceutical Engineering magazine on applying GAMP concepts to machine learning (ispe.org) (ispe.org).

What is a reasonable revalidation strategy after an AI model change? Tier the response by risk: for low-risk functions, a lighter regression check and updated audit-trail entry may suffice; for higher-risk functions, a fuller impact assessment, pre-specified acceptance criteria, item-level regression testing, and a documented rollback plan should all be completed before the new version is used in production, consistent with the risk-scaling principle in ICH Q9(R1) (database.ich.org).

Conclusion

Regression testing an AI model change inside a validated workflow is not, at its core, a new regulatory category. Applicable change-control expectations for a model change should be assessed against the governing framework and system scope. What has genuinely changed is the operational tempo: model providers now version, deprecate, and replace models on cycles as short as two weeks for preview releases and as long as eighteen months for some general-availability models, and every one of those cycles is a potential trigger for a validated workflow's own change-control process. Organizations that pin dated model identifiers, monitor vendor deprecation policies as a standing responsibility, build item-level regression test suites against representative production data, and pre-plan rollback before a swap rather than after a failure are applying recognized engineering practices. Whether a control is a regulatory obligation depends on the applicable framework and the relevant regulated activity, system, and record scope. The frameworks with the clearest, most explicit answers, FDA's PCCP and GMLP guidance, apply narrowly to authorized AI-enabled medical devices today, which leaves most life sciences organizations extending general GxP change-control language to LLM-specific changes by analogy rather than by direct rule. That gap is likely to narrow as industry bodies such as ISPE publish more AI-specific guidance, but until it does, the risk-based, documented, test-before-you-switch discipline described throughout this report remains the most defensible available practice.
