



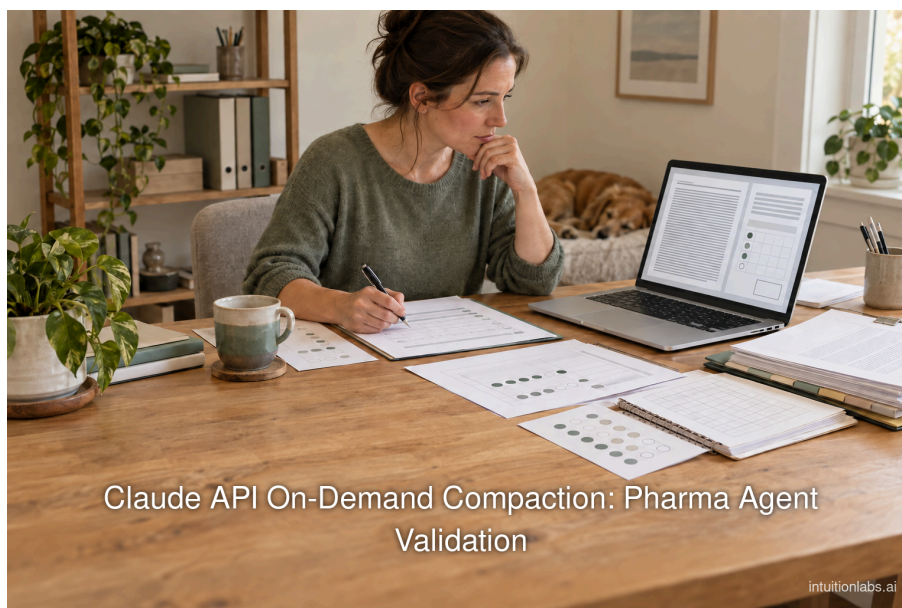
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude API On-Demand Compaction: Pharma Agent Validation

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-10-05 · claude api on-demand compaction · claude context compaction · anthropic · pharmaceutical ai agents · agent validation · context retention · tool use testing · audit trails · llm context management

Original article: <https://intuitionlabs.ai/articles/claude-api-compaction-pharma-agent-validation>



In this article

1. [Executive Summary](#)
 2. [Introduction and Background](#)
 3. [Key Changes](#)
 4. [Implementation Considerations and Process Changes](#)
 5. [Critical-State Inventory for Pharmaceutical Agents](#)
 6. [Regression Harness and Context-Retention Testing](#)
 7. [Tool Use, Failure Modes, and Recovery](#)
 8. [Audit Evidence and Operational Ownership](#)
 9. [Data Analysis and Evidence](#)
 10. [Implications and Future Directions](#)
 11. [Conclusion](#)
-

Executive Summary

Claude API on-demand compaction gives developers control over when a conversation is summarized through the application programming interface (API). Anthropic introduced the beta on **September 14, 2026**; the release described a signed block, background execution, and retention of recent turns.

(platform.claude.com) This report's central recommendation is to treat summary replacement as a **lossy state transformation** with explicit acceptance tests. A theoretical compaction paper examines selection and bounded summarization, but models a single invocation; it provides evaluation framing rather than a guarantee for a production agent. (arxiv.org) (www.w3.org)

The **compact-2026-09-04** header applies to generation and later requests carrying the block.

(platform.claude.com) Current documentation, checked for this report on **October 5, 2026**, lists the Claude API, Claude Platform on AWS, Google Cloud, and Microsoft Foundry in beta. Amazon Bedrock is absent from that on-demand support list. (platform.claude.com) Kept thinking has additional history and model constraints, so successful summary generation alone is an incomplete acceptance check. (platform.claude.com) A signature establishes an API acceptance mechanism; semantic completeness must be evaluated separately against authoritative state and intended use. The applicable record requirements must first be established rather than assumed. (www.fda.gov)

For pharmaceutical workflows, the proposed golden set covers **case identity, adverse-event facts, evidence provenance, protocol constraints, knowledge updates, and pending actions**. It tests retrieval and decisions before and after replacement, including repeated compactions and concurrent tool activity. Long-context research supports position-sensitive testing; long-term memory evaluation also covers temporal reasoning, updates, and abstention. These findings guide test design and do not measure this compaction feature. (aclanthology.org) (openreview.net) Regulatory scope remains workflow-specific: the Food and Drug Administration (FDA) recommends **documented, risk-based validation**, while closed-system controls include accurate and complete record copies. (www.fda.gov) (www.ecfr.gov)

The economic decision requires [actual request usage](#). Current **Claude Sonnet 5.5** list prices are **\$2 per million input tokens** and **\$10 per million output tokens**. ([platform.claude.com](#)) Token counting provides estimates for planning. ([platform.claude.com](#)) This report performs no paid API experiment and reports no savings percentage or preservation rate. Adoption should require a passing semantic regression suite, recoverable original history, controlled tool permissions, and an auditable replacement process. These are proposed engineering controls, supported by provenance standards and downstream authorization guidance, rather than vendor or regulatory certification. ([www.w3.org](#)) ([genai.owasp.org](#))

Introduction and Background

Long-running pharmaceutical agents accumulate more than conversational prose. A useful validation boundary is the state needed to continue the task correctly: the governing protocol, authoritative case facts, cited evidence, unresolved questions, and permissions for future actions. This report proposes making that boundary explicit before implementing context compaction. The rationale is consistent with the World Wide Web Consortium (**W3C**) definition of provenance, which connects data to the entities, activities, and people involved in producing it. ([www.w3.org](#))

Context compaction replaces older conversation turns with a summary. ([platform.claude.com](#)) It therefore changes the representation supplied to the model. The original transcript may remain in application storage, but the compaction theory paper explains that discarded information no longer reaches the model unless subsequent inputs restore it. This distinction motivates separate controls for record preservation and operational recall. ([arxiv.org](#)) A recovery archive is useful only if the application can retrieve the relevant material when needed.

The audience is architects, machine-learning engineers, validation teams, and owners of safety, regulatory, clinical, and scientific agents. **Good practice**, commonly abbreviated **GxP**, is used here as a collective label for regulated practice domains; the Medicines and Healthcare products Regulatory Agency (**MHRA**) guidance covers manufacturing, distribution, and pharmacovigilance among its sectors. ([www.gov.uk](#)) Applicability must be decided for the actual workflow. [FDA's Part 11 guidance](#) recommends documenting those scope decisions. ([www.fda.gov](#))

IntuitionLabs' adjacent advisory perspective is reflected in its stated emphasis on connecting artificial intelligence (AI) to authoritative enterprise sources with identity, permissions, retrieval, citations, and evaluation. ([intuitionlabs.ai](#)) The technical analysis below consequently focuses on demonstrable continuity of work. The Organisation for Economic Co-operation and Development (**OECD**) similarly frames [computerized-system validation](#) around risk assessment and a lifecycle approach. ([www.oecd.org](#)) ([www.oecd.org](#)) The proposed harness is a concrete implementation design, not a claim that these sources prescribe a particular AI architecture.

The regulatory sources serve different purposes. FDA's scope guidance describes recommendations, while **Title 21 of the Code of Federal Regulations (CFR), section 11.10**, supplies applicable closed-system controls; OECD guidance emphasizes scalable risk assessment. These should be read according to their scope, not combined into an automatic approval of an agent implementation. ([www.fda.gov](#)) ([www.ecfr.gov](#)) ([www.oecd.org](#))

Key Changes

Developer-controlled timing and platform scope

Threshold compaction runs within the request that reaches its trigger; on-demand compaction lets the developer choose the moment. (platform.claude.com) *Table 1* compares the operational choices. Its validation column contains proposed checks rather than documented performance promises.

Pattern	Documented behavior	Proposed acceptance check
Threshold	Summarization occurs in the threshold-reaching request. (platform.claude.com)	Reproduce the boundary and inspect the next task decision.
On demand	Developer chooses when to request a signed summary. (platform.claude.com)	Confirm the selected prefix matches the archived snapshot.
Background	Full-history work continues while compaction runs. (platform.claude.com)	Reject a result if its snapshot lineage is stale.
Keep recent turns	Application chooses the cut; tool calls and results stay together. (platform.claude.com)	Compare retained messages exactly and test pending work.

The choice should follow the task's state transitions. This report recommends a boundary after evidence has been committed and before a new decision phase begins. That is an application policy, not a vendor timing requirement. Modeling replacement as derivation makes the old state, transformation, and new state identifiable. (www.w3.org) Atomic transaction mechanisms offer a useful implementation pattern for committing that transition. (www.postgresql.org)

The current supported model families are **Fable 5/5.1**, **Mythos 5/5.1/Preview**, **Opus 4.6/4.7/4.8/5/5.5**, and **Sonnet 4.6/5/5.5**. (platform.claude.com) Teams should pin the selected model and platform in qualification evidence; availability on one named service should not be generalized to another service with a similar cloud label.

Signed-block lifecycle

Replace the summarized prefix with exactly one unchanged signed block at the start of continuation history. (platform.claude.com) Signing and API content checks should be treated as structural acceptance controls. The separate question is whether the transformed state still supports the intended task. Closed-system validation language addresses accuracy, reliability, and consistent intended performance, providing a broader assessment boundary. (www.ecfr.gov)

This report proposes three independent verdicts: **transport accepted**, **critical state preserved**, and **authorized continuation correct**. A passing transport verdict must never substitute for the other verdicts. Structure can be checked with schemas, but object fields are optional unless explicitly required. (json-schema.org) Provenance can establish which inputs produced a state without proving that every clinically relevant distinction was retained. (www.w3.org)

Constraints that affect implementation

On-demand compaction rejects `context_management`, `stop_sequences`, structured-output format, forced `tool_choice`, and `output_config.task_budget.remaining`. Its `max_tokens` budget includes preceding thinking. (platform.claude.com) (platform.claude.com) Keep these constraints in a versioned request builder, separate from application acceptance criteria. The regression fixture should capture the exact request configuration and expected response handling so parameter changes become visible changes to intended behavior. Shared setup and cleanup are standard fixture capabilities. (docs.python.org)

For each acceptance case, record the transformed inputs, required response, observed result, source-state derivation, and schema version. (www.w3.org) (docs.python.org) (json-schema.org)

Implementation Considerations and Process Changes

Snapshot, continue, and commit

The documented background pattern appends turns while preserving existing history and avoiding another compaction. When the block arrives, only the submitted prefix is replaced; subsequently appended turns stay after it. (platform.claude.com) (platform.claude.com) The application should represent this as a state machine with **full history**, **compaction pending**, **candidate ready**, **validated replacement**, and **recovery** states. These labels are the report's proposed implementation vocabulary.

The following transition discipline is recommended:

- **Snapshot:** Persist the exact submitted prefix, its boundary, and the active configuration.
- **Continue:** Append new work under the same history lineage while the candidate is pending.
- **Check:** Associate the returned candidate with the snapshot that generated it.
- **Validate:** Run critical-state probes before allowing consequential continuation.
- **Commit:** Publish the replacement and its lineage reference together.
- **Recover:** Keep the previous valid branch available if acceptance fails.

Use an application-level version precondition on the commit. A transaction can make several database updates atomic and hide intermediate states from concurrent transactions. It does not supply the application's semantic check or identify a stale compaction candidate automatically. (www.postgresql.org) Connect the asynchronous request and the foreground work through trace correlation; OpenTelemetry defines identifiers that group spans across processes. (opentelemetry.io) The durable state record should remain the source of truth, with telemetry assisting diagnosis.

Keep-tail boundaries and thinking

Kept turns must be sent unchanged, including their thinking blocks. Tool calls and results must remain on the same side of the cut. (platform.claude.com) A practical qualification case should place an unresolved tool action near the boundary and verify that the application either delays the cut or retains a coherent interaction. Downstream permission checks should still evaluate every action independently. (genai.owasp.org)

For preserved thinking, every intervening compaction must use a preserved-thinking model. Kept messages must immediately follow the summarized messages unchanged, their first role must differ from the preceding role, and the system and non-deferred tools must match across the relevant requests. (platform.claude.com) Configuration equality should therefore be an explicit precondition in the proposed state machine. A schema can enforce the presence of configuration identifiers but cannot establish the truth of their contents. (json-schema.org)

Use one binding-check model throughout a thinking-enabled compatibility fixture. Retain at least one thinking block; send both compact-2026-09-04 and thinking-binding-controls-2026-08-01; and set `thinking.block_binding.prefix_mismatch_behavior` to "error". Continue with the signed block, unchanged kept turns, then a new user message. Require HTTP **200** with empty `input_transformations` only under these preconditions. For a negative control, alter a retained message preceding the thinking block, leaving both the thinking and signed compaction blocks unchanged; require HTTP **400** identifying a conversation-binding mismatch. Reject compatibility if either control fails, then run behavioral probes separately. ([Anthropic documentation](#))

Mid-conversation instructions

Instructions in summarized role: "system" messages stop applying when the block replaces them, even if their content is remembered in the summary. ([Anthropic documentation](#)) Add a protocol-constraint fixture with such a governing instruction. Restate any still-required instruction in a role: "system" message directly after the first new user turn following the kept turns, and retain it in later history. Leave the kept turns unchanged; placing the restatement between the block and kept turns breaks their thinking. ([Anthropic documentation](#)) Verify the expected authorized decision and repeat the positive thinking check. Remembering instruction content is separate from preserving its instruction role.

Evidence attachments and configuration control

Images, documents, container uploads, and fetched URLs inside replaced messages are absent from continuation history. (platform.claude.com) This report recommends retaining resolvable evidence references outside the summary and testing rehydration explicitly. W3C provenance provides a vocabulary for relating the derived state to its source entities. (www.w3.org) For clinical-trial migration, the European Medicines Agency (**EMA**) recommends keeping data, contextual information, and audit trails together. That recommendation informs the archive design; it is not an approval of summary replacement. (www.ema.europa.eu)

Treat the chosen boundary as a reviewable configuration artifact. Record its source-state relationship and correlate asynchronous activity across processes. When accepting a candidate requires several state updates, commit them atomically so readers cannot observe an incomplete replacement. These are proposed application controls built on documented infrastructure capabilities. (www.w3.org) (opentelemetry.io) (www.postgresql.org)

Critical-State Inventory for Pharmaceutical Agents

Define preservation before measuring it

The proposed **critical-state contract** is a typed, versioned inventory of what subsequent decisions require. It should distinguish exact-match identifiers, meaning-sensitive facts, documentary evidence, temporal relationships, and action status. MHRA describes metadata as supplying attributes and context for other data, making it a useful basis for deciding what a prose summary alone cannot represent reliably.

(assets.publishing.service.gov.uk) FDA's guidance also emphasizes copies that preserve record content and meaning. (www.fda.gov)

Table 2 defines a suggested golden set. **Before** and **after** columns describe measurements to collect, not experimental results. Each row must be instantiated with approved fixtures appropriate to the agent's intended use.

Critical state	Before compaction	After compaction	Proposed release criterion
Case and document identifiers	Exact authoritative values	Exact values returned by probes	No changed or missing required identifier. (json-schema.org)
Adverse-event facts	Facts, uncertainty, source associations	Matched facts and qualifiers	No unsupported change to required meaning. (www.fda.gov)
Citations and evidence	Source identifier, version, location	Resolvable matching evidence	No invented or misattributed supporting source. (www.w3.org)
Protocol constraints	Applicable version and conditions	Constraint-aware task decision	Same authorized decision under unchanged facts. (www.ecfr.gov)
Temporal updates	Ordered corrections and superseded facts	Current value plus update history	Correct latest state without erasing lineage. (openreview.net)
Pending actions	Intent, status, tool operation identifier	Correct next action and status	No duplicate or unauthorized operation. (genai.owasp.org)
Unknowns and conflicts	Explicit unanswered questions	Correct uncertainty or abstention	No fabricated resolution. (openreview.net)

A fluent answer is insufficient evidence of preservation. The table deliberately includes attribution, qualifiers, and executable obligations, because the proposed acceptance boundary is task continuity. LongMemEval's evaluation categories include extraction, temporal reasoning, updates, and abstention, supporting this broader test design. (openreview.net) The International Council for Harmonisation (**ICH**) recommends reconciliation during data exchange or migration to avoid loss and unintended modifications; this is relevant context for controlled state replacement. (database.ich.org)

Separate operational memory from regulated records

Classify each inventory field by authoritative location, intended use, responsible owner, and retention policy. A working summary can be disposable operational memory while its source remains an official record. Where the agent output contributes to a regulated record, the applicable record rules need a documented

assessment. (www.fda.gov) Applicable closed-system controls include accurate, ready retrieval throughout retention and preservation of prior information when records change. (www.ecfr.gov)

For each field, specify:

- **Authority:** The source that determines the accepted value.
- **Identity:** The immutable identifier and relevant version.
- **Meaning:** Qualifiers, units, negation, uncertainty, and relationships.
- **Access:** The permission needed to retrieve or act on it.
- **Continuity:** The next task that depends on the field.
- **Recovery:** The route to reconstruct state when a probe fails.

This inventory should be reviewed as an intended-use specification. OECD's risk-centered lifecycle approach supports scaling validation effort to the significance of the system. (www.oecd.org) The contract should explicitly state which fields require exact equality and which need adjudicated semantic equivalence, making a passing result reviewable rather than impressionistic.

For changing fields, include the current value, superseded value, update order, and authoritative source in the fixture; mark required fields in its schema. (openreview.net) (www.w3.org) (json-schema.org)

Regression Harness and Context-Retention Testing

Build paired tests against authoritative state

The proposed harness forks an approved transcript into an uncompact branch and a compacted branch. Both receive identical continuation tasks and tool fixtures. Score each branch against independent authoritative labels; comparing branches alone could accept an error shared by both. The National Institute of Standards and Technology (NIST) Generative AI Profile recommends ground-truth comparison and empirical assessment of capability claims. (nvlpubs.nist.gov) Test cases should specify input and expected behavior rather than merely confirm that a function returned a response. (docs.python.org)

Qualification should include these test families:

- **Fact recall:** Probe every required identifier and meaning-sensitive field.
- **Citation preservation:** Check source identity, version, location, and supporting passage.
- **Decision continuity:** Compare the next task outcome against the approved answer.
- **Position variation:** Move critical evidence among early, middle, and recent turns. (aclanthology.org)
- **Update handling:** Insert later corrections and test which value governs. (openreview.net)
- **Repeated replacement:** Evaluate several successive summaries, not only the first. (arxiv.org)
- **Tool sequencing:** Exercise action status and permission checks. (genai.owasp.org)
- **Boundary variation:** Repeat cases with different retained-tail cuts and history lengths. (docs.python.org)

Treat the uncompressed branch as a baseline, not an oracle. Lost in the Middle shows that relevant-information position can change performance, including in long-context models studied in that work. (aclanthology.org) A compacted branch may improve one retrieval task while degrading another. Report results

by risk category so gains in ordinary recall cannot mask changed case identity or lost constraints.

Make the oracle explicit

Use exact equality for identifiers, typed comparisons for structured fields, and source-grounded adjudication for paraphrased meaning. JavaScript Object Notation (JSON) Schema should mark required properties explicitly and decide whether extra properties are allowed. (json-schema.org) (json-schema.org) Check each asserted onset date, causal qualifier, and source relationship against the independent expected answer.

Partition tests by **indexing**, **retrieval**, and **reading**, the memory-system stages described in the LongMemEval paper. (proceedings.iclr.cc) If an archived source cannot be found, that is a retrieval problem; if it is retrieved but ignored, that is a reading problem. Use this separation to avoid blaming summary quality for every downstream deviation. An evidence lineage record links the relevant source entity to the derived state and observed answer. (www.w3.org)

A corrected-case fixture (Hypothetical Example)

Construct a synthetic safety case with an initial reported fact, a later correction, an unresolved follow-up, and a draft action that has not been executed. Before compaction, ask which fact is current and which action is pending. After compaction, repeat those probes and request the next workflow step. The proposed expected result preserves the correction, cites its source, and keeps the action unexecuted until authorized. Update handling and abstention are independently recognized evaluation dimensions; downstream authorization remains a separate control. (openreview.net) (genai.owasp.org)

Vary the correction's position and the cut boundary while reusing the same fixture. Shared preparation and cleanup make this practical, but the test must retain an independent expected answer. (docs.python.org) No preservation percentage is implied by this example. Release thresholds should be approved through a documented risk assessment and should explicitly identify which failures prevent deployment. (www.fda.gov)

Include retrieval tasks that require combining evidence, rather than testing isolated strings exclusively. RULER includes multi-hop and aggregation tasks, while the staged memory model distinguishes finding evidence from reading it. Expected outcomes should still be expressed as concrete test assertions for the actual agent. (arxiv.org) (proceedings.iclr.cc) (docs.python.org)

Tool Use, Failure Modes, and Recovery

Preserve action semantics across replacement

The proposed tool-use suite distinguishes an intended call, an issued call, a completed result, and a committed business action. Each transition should have durable state outside the summary. Authorization belongs in downstream systems rather than in the model's recollection of a permission grant. (genai.owasp.org) Applicable closed-system controls also include authority checks. (www.ecfr.gov)

Use separate fixtures for a delayed result, a retry after an uncertain response, a cancelled action, and a resumed session. Trace identifiers can connect the foreground task to the asynchronous summarization activity. (opentelemetry.io) The oracle should verify the operation's final state as well as the assistant's

description of it. A correct narrative cannot substitute for a database or destination-system check.

Table 3 separates documented API outcomes from proposed application responses. The semantic test row is an application failure condition, not an Anthropic error code.

Observed condition	Meaning	Proposed response
400 compaction_block_misplaced, compaction_signature_invalid, or compaction_content_mismatch	Placement or block validation failed. (platform.claude.com)	Repair serialization; retain the previous valid history.
529 compaction_unavailable	Transient summarization problem. (platform.claude.com)	Bounded retry with backoff; continue only within safe operating limits.
200 without a usable block	Summarization did not necessarily produce a candidate. (platform.claude.com)	Inspect stop_reason; reject unusable candidates, raise the budget for max_tokens, or shorten submitted input for model_context_window_exceeded.
Continuation thinking mismatch	Compatibility may fail or drop thinking. (platform.claude.com)	Reject the branch until the continuation policy passes.
Critical-state probe fails	Application acceptance criterion failed. (docs.python.org)	Quarantine candidate, restore the valid branch, review the deviation.

Retry policy should follow the failure class. Repeating an invalid block does not change its validity, while a transient availability response may justify a bounded retry. Recovery should preserve enough context to explain which branch was accepted and why. Provenance records can relate the candidate and retained branch to the same source history. (www.w3.org) Database rollback is a useful local implementation mechanism, although it does not reverse an external business action. (www.postgresql.org)

Test recovery as deliberately as success

The recovery plan should contain:

- **Retry limits:** Bound attempts and elapsed time according to workflow needs.
- **Candidate isolation:** Prevent an unaccepted summary from becoming active state.
- **Branch restoration:** Reload the last validated history and its configuration.
- **Operation reconciliation:** Check downstream action status before reissuing work.
- **Escalation:** Route unresolved consequential decisions to the designated reviewer.

These are proposed operational controls. OWASP recommends approval for high-impact actions, and ICH recommends contingency procedures for essential data. (genai.owasp.org) (database.ich.org) Exercise the recovery path with deterministic fixtures rather than waiting for production exceptions. Test preparation should include the state needed to reproduce each failure and cleanup that restores the environment. (docs.python.org)

Recovery evidence should identify the interrupted transformation, the retained branch, and the operation's authorized status. Database transactions support local rollback, but model-side narration must not determine access. Keep the source-state relationship available to the reviewer when deciding how to resume.

(www.postgresql.org) (genai.owasp.org) (www.w3.org)

Audit Evidence and Operational Ownership

Record the transformation and the decision

The proposed **compaction audit packet** should make the replacement reconstructable. W3C PROV models provenance around data entities, activities, and responsible participants, while its data model defines derivation as transformation between entities. (www.w3.org) (www.w3.org) This offers a useful organizing scheme without establishing regulated electronic-signature compliance.

For each accepted or rejected candidate, capture:

- **Source lineage:** Transcript snapshot identity, boundary, and authoritative references.
- **Configuration:** Model, platform, request builder, system, tools, and test-contract versions.
- **Candidate evidence:** Returned block, response status, request identifier, and usage record.
- **Compatibility verdict:** Serialization checks and continuation-thinking observations.
- **Semantic verdict:** Probe outputs, expected labels, deviations, and adjudication.
- **Operational outcome:** Accepted branch, reviewer decision, recovery action, and downstream state.

Store the block without editing its contents, but keep the explanatory packet separately. The application can link related spans across traces using OpenTelemetry links. (opentelemetry.io) Durable evidence should retain the relationship between inputs, transformation, and continuation; traces alone are not a substitute for the record archive. WHO guidance recommends retaining electronic records with associated metadata, including audit trails and electronic signatures. (cdn.who.int)

Establish scope, access, and review ownership

The validation owner should approve intended use and release criteria; the application owner should control request changes and recovery; the workflow owner should adjudicate domain semantics. These are suggested responsibilities. OECD's lifecycle framing supports treating validation and operation as connected activities rather than a one-time integration check. (www.oecd.org) NIST's AI Risk Management Framework is voluntary, so using its terminology should not be presented as a certification. (www.nist.gov)

Where closed-system requirements apply, prior recorded information must not be obscured by changes, and audit documentation must be retained with the associated records for the required period. (www.ecfr.gov) EMA's guidance similarly recommends preserving contextual information and audit trails during migration. (www.ema.europa.eu) Decide whether the working summary belongs in the regulated record set and document the rationale, rather than assigning a universal retention duration to every conversation.

Access to the archive should preserve the same practical permission boundary as access to the source. Downstream authorization guidance supports enforcing this outside the language model. (genai.owasp.org) Avoid logging sensitive content indiscriminately; choose the minimum evidence needed for reproducibility and

review under the workflow's approved information controls. Provenance relationships can refer to controlled entities without requiring every telemetry event to duplicate their contents. (www.w3.org)

Assign each evidence item to the durable record packet or operational telemetry, and record its source reference, trace link, and retention owner. (www.w3.org) (opentelemetry.io) (www.ecfr.gov)

Data Analysis and Evidence

What existing research measures

The relevant public evidence concerns evaluation design rather than validated pharmaceutical compaction performance. **RULER** evaluates **17 long-context models** using **13 representative tasks**, extending retrieval with multi-hop tracing and aggregation. (arxiv.org) **LongMemEval** contains **500 curated questions** and evaluates extraction, multi-session reasoning, temporal reasoning, updates, and abstention. (openreview.net) Lost in the Middle studies multi-document question answering and key-value retrieval, with position-dependent performance. (aclanthology.org)

The **August 2, 2026** version of Context Compaction Theory examines a single compaction invocation and distinguishes retained subsets from bounded generated summaries. (arxiv.org) These studies should not be pooled into a claimed success rate for Anthropic's on-demand feature. Their tasks, models, and methods differ. Their practical contribution here is a test taxonomy: probe retrieval, compositional reasoning, changing facts, uncertainty, and successive state transformations separately. (arxiv.org) (proceedings.iclr.cc)

Measure retention and task continuity

The proposed report card should calculate **required-fact recall** as correctly recovered required facts divided by applicable required facts. Calculate **citation preservation** as correctly associated, resolvable evidence references divided by required references. Measure **decision accuracy**, **pending-action continuity**, and **unsupported additions** separately. These are analyst-defined measures, not published results. Independent expected labels and concrete assertions make their denominators auditable. (docs.python.org) (json-schema.org)

Report before and after values for every golden-set row, with the number of applicable cases, excluded cases, and adjudicated disagreements. Split the results by workflow, risk category, history position, tail boundary, and compaction count. Position-sensitive tests are motivated by the long-context literature; indexing, retrieval, and reading should be diagnosed independently. (aclanthology.org) (proceedings.iclr.cc) State the conditions under which the evidence applies instead of reporting one blended score as a universal preservation guarantee.

Calculate economics from actual API usage

For an uncached continuation, the proposed input-only calculation is: **uncompacted input tokens minus compacted input tokens, multiplied by the applicable input price per token**. Sonnet 5.5's current input rate is **\$2 per million tokens**. (platform.claude.com) This calculation is a reporting formula, not a measured saving. It excludes summarization and output costs and should be used only with actual usage from comparable continuations.

On-demand generation is billed; its cost is in usage . iterations, while top-level input and output tokens are zero. ([platform.claude.com](#)) Cached input totals combine input_tokens, cache_creation_input_tokens, and cache_read_input_tokens. ([platform.claude.com](#)) Sonnet 5.5 cache rates are **\$2.50** for five-minute writes, **\$4** for one-hour writes, and **\$0.20** for reads per million tokens. ([platform.claude.com](#)) Account for each category at its own rate, plus output, retries, and additional retrieval work.

Token counting is an estimate, is free within rate limits, and does not simulate cache behavior. ([platform.claude.com](#)) Therefore a smaller estimated prompt is not a measured billing result. The economic experiment should compare complete sessions under the same task and cache policy, then relate cost to the already approved semantic outcomes. OECD's risk-based scaling principle supports prioritizing acceptable performance over an isolated efficiency figure. ([www.oecd.org](#))

A benchmark description is useful only within its method. RULER's synthetic tasks broaden retrieval testing; LongMemEval's curated conversational questions cover changing and temporal information. They motivate different fixture families, not a conversion factor from context length to pharmaceutical correctness. This report deliberately keeps those populations separate from the proposed workflow tests. ([arxiv.org](#)) ([openreview.net](#))

To interpret a failed probe, retain the expected answer, relevant source, actual output, and failure classification. A specific input-response assertion and an explicit state schema support reproducibility; the indexing, retrieval, and reading distinction supports diagnosis. This proposed report card should accompany cost measurements rather than be replaced by them. ([docs.python.org](#)) ([json-schema.org](#)) ([proceedings.iclr.cc](#))

Implications and Future Directions

The recommended adoption decision is conditional: proceed only when the state contract, paired regression, and recovery evidence support the specific workflow. The following decision sequence is proposed, rather than presented as a regulatory checklist:

- **Scope the work:** Establish whether agent state or outputs contribute to required records. ([www.fda.gov](#))
- **Check compatibility:** Qualify the selected platform, model, and request builder.
- **Define critical state:** Approve authoritative labels and semantic acceptance rules. ([www.ecfr.gov](#))
- **Test the transformation:** Run paired recall, citation, decision, and boundary cases. ([docs.python.org](#))
- **Test continuation:** Verify authorization and downstream operation state. ([genai.owasp.org](#))
- **Prove recovery:** Reconstruct the valid branch and reconcile pending work. ([www.postgresql.org](#))
- **Review change:** Requalify affected cases when configuration or intended use changes. ([www.oecd.org](#))

For a scientific reading assistant, the contract might emphasize source-grounded retrieval; for a workflow agent, operation status and permission can be decisive. Those are intended-use choices, not measured product distinctions. LongMemEval's staged memory-system analysis provides a useful way to separate source discovery from interpretation when diagnosing these different workloads. ([proceedings.iclr.cc](#))

Future vendor changes should be incorporated through a controlled delta review. Preserve the earlier evidence, identify altered interfaces or behavior, and select affected fixtures. The theory paper's single-invocation scope also leaves repeated compaction as a separate empirical question for the application. ([arxiv.org](#)) No inferred guarantee should bridge that gap.

IntuitionLabs' stated measurement perspective includes quality, risk signals, reliability, and support burden before expansion. (intuitionlabs.ai) Applied here, that means scaling only after both task performance and operational ownership are demonstrated. NIST's voluntary risk framework offers broader governance context; it does not replace workflow-specific qualification. (www.nist.gov)

Document the deployment organization's record scope, acceptance thresholds, reviewer authority, and lifecycle review owner. (www.nist.gov) (www.oecd.org)

Conclusion

Adopt Claude API on-demand compaction when the qualification package supports the intended workflow: an approved critical-state inventory, independent expected answers, paired before-and-after tests, passing continuation controls, and exercised recovery. Record which state was submitted, which continuation was tested, which critical facts survived, who reviewed the result, and how exceptions were resolved. ([www.w3.org](https://www.w3.org/docs.python.org)) (genai.owasp.org) (www.w3.org)

Assess the workflow's record scope and risk before release. Where applicable, verify accurate copies, recoverable history, and reviewable changes. (www.ecfr.gov) Base the economic decision on actual session usage and category-specific costs, reported alongside the test outcomes.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://platform.claude.com/docs/en/release-notes/overview>
2. <https://arxiv.org/html/2608.01326v1>
3. <https://www.w3.org/TR/prov-dm/>
4. <https://platform.claude.com/docs/en/build-with-claude/compaction-on-demand>
5. <https://platform.claude.com/docs/en/build-with-claude/compaction-background>
6. <https://platform.claude.com/docs/en/build-with-claude/compaction-thinking-blocks>
7. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application>
8. <https://aclanthology.org/2024.tacl-1.9/>
9. <https://openreview.net/forum?id=pZiyCaVuti>
10. <https://intuitionlabs.ai/articles/risk-based-ai-validation-pharma-ich-q9>
11. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10>
12. <https://intuitionlabs.ai/articles/token-optimization-chatgpt-claude-costs>
13. <https://platform.claude.com/docs/en/about-claude/pricing>
14. <https://platform.claude.com/docs/en/build-with-claude/token-counting>
15. <https://www.w3.org/TR/prov-overview/>
16. <https://genai.owasp.org/llmrisk/llm062025-excessive-agency/>
17. <https://platform.claude.com/docs/en/build-with-claude/compaction>

18. <https://www.gov.uk/government/publications/guidance-on-gxp-data-integrity>
19. <https://intuitionlabs.ai/articles/21-cfr-part-11-compliance-requirements>
20. <https://intuitionlabs.ai/>
21. <https://intuitionlabs.ai/articles/csv-pharmaceutical-biotech-compliance>
22. https://www.oecd.org/en/publications/application-of-glp-principles-to-computerised-systems_e3c9cda5-en.html
23. <https://platform.claude.com/docs/en/build-with-claude/compaction-keep-recent-turns>
24. <https://www.postgresql.org/docs/current/tutorial-transactions.html>
25. <https://json-schema.org/understanding-json-schema/reference/object>
26. <https://docs.python.org/3/library/unittest.html>
27. <https://opentelemetry.io/docs/specs/otel/overview/>
28. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-electronic-data-clinical-trials_en.pdf
29. https://assets.publishing.service.gov.uk/media/5aa2b9ede5274a3e391e37f3/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf
30. https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf
31. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
32. https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html
33. <https://arxiv.org/abs/2404.06654>
34. https://cdn.who.int/media/docs/default-source/medicines/norms-and-standards/guidelines/inspections/trs1033-annex4-guideline-on-data-integrity.pdf?sfvrsn=6218a4e6_4
35. <https://www.nist.gov/itl/ai-risk-management-framework>
36. <https://platform.claude.com/docs/en/build-with-claude/prompt-caching>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.