



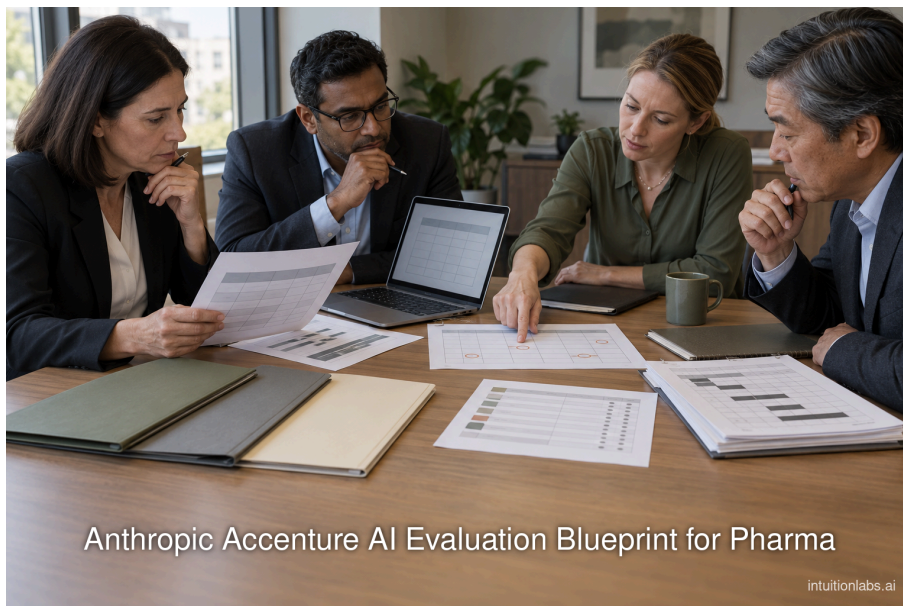
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Anthropic Accenture AI Evaluation Blueprint for Pharma

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 · 2026-10-13 · anthropic accenture ai evaluation for pharma · embedded ai evaluation · independent ai assurance · pharmaceutical ai validation · claude life sciences · third party ai assurance · ai governance pharma · continuous ai evaluation

Original article: <https://intuitionlabs.ai/articles/anthropic-accenture-ai-evaluation-blueprint-pharma>



In this article

1. Executive Summary
 2. Introduction and Background
 3. What Embedded Evaluation Adds to AI Assurance
 4. The Three Assurance Layers in Pharma
 5. Evidence Access Without Exposing Confidential R&D
 6. Independence, Reporting, and Remediation Ownership
 7. Continuous Evaluation and Procurement Decisions
 8. Data Analysis and Evidence
 9. Implications and Future Directions
 10. Frequently Asked Questions (FAQs)
 11. Conclusion
-

Executive Summary

Pharma buyers should require **assurance at the model, application, and context-of-use layers**, with explicit evidence rights and decision owners. Anthropic's public system cards supply model-level documentation; deployment-context testing and sponsor responsibility require additional local evidence. An embedded evaluator's access to a frontier-model developer does not by itself establish that a sponsor's configured workflow is fit for its intended decision. (www.anthropic.com) (airc.nist.gov) (www.ema.europa.eu)

The **September 18, 2026** Anthropic and Accenture announcement is a concrete trigger for this procurement discussion. (newsroom.accenture.com) Accenture describes a team working alongside Anthropic's internal teams and safety partners, with each company expecting to invest **at least \$1 billion over five years** in artificial intelligence (AI) safety. (newsroom.accenture.com) This is an investment expectation, not a published customer price or completed validation result. The arrangement should be assessed through its disclosed operating terms, rather than treated as a ready-made pharma assurance standard.

As of **October 13, 2026**, the United States Food and Drug Administration (FDA) still labels its **January 2025** AI credibility guidance draft and not for implementation. (www.fda.gov) Its scope is AI-produced information supporting regulatory decisions about drug safety, effectiveness, or quality, not every operational AI workflow. (www.fda.gov) The National Institute of Standards and Technology (NIST) **AI 200-2 TEVV-Athlon**, a Test, Evaluation, Verification, and Validation (TEVV) framework, remains an **initial public draft**, with its **October 6, 2026** comment deadline already passed. (www.nist.gov) Its **four-stage** approach can inform evaluation planning. (nvlpubs.nist.gov) Neither document establishes a requirement to appoint embedded evaluators.

This report proposes a buyer-side design: compare snapshot, recurring, and embedded assessment; specify evidence access without unrestricted research and development (R&D) disclosure; document funding, scope, reporting, and conflict controls; and reopen approval after material changes. The National Telecommunications and Information Administration (NTIA) frames evaluator access alongside privacy, security, and intellectual-property protection, while the Open Worldwide Application Security Project (OWASP) (owasp.org) recommends retesting changes to prompts, tools, retrieval, policies, and providers. (www.ntia.gov) (cheatsheetseries.owasp.org) A practical decision record should distinguish tested coverage from acceptable

residual risk, identify unknown independence terms, and assign remediation to the party controlling the affected component. Confidentiality can be strengthened through project segregation and transcript access controls, as illustrated by METR's public description of its own practices. (metr.org)

Introduction and Background

The procurement question behind **Anthropic Accenture AI evaluation for pharma** is whether an evaluator can obtain enough evidence to challenge a model supplier's conclusions, while the sponsor retains control over confidential research and regulated records. The answer depends on the assurance claim being purchased. United Kingdom (UK) government guidance defines assurance through measurement, evaluation, and communication; it also recommends combining techniques according to the deployment context.

(www.gov.uk) (www.gov.uk)

The concrete trigger is the **September 18, 2026** announcement of an Anthropic and Accenture partnership led by **Faculty**, Accenture's specialist AI business, covering model evaluation, red teaming, alignment assessments, and safeguards. (www.anthropic.com) The parallel announcement says each company expects to invest **at least \$1 billion over five years** in AI safety. This is forward-looking investment language, not a disclosed evaluation contract price, sponsor fee, or completed assurance result. (newsroom.accenture.com)

This report separates **observations** in those publications from a **proposed buyer framework**. Employee-comparable internal access could help evaluators understand development decisions, but Anthropic says operating standards remain unsettled. The announcement also describes a nonexclusive arrangement. These statements establish an emerging approach, rather than a standardized accreditation scheme.

(www.anthropic.com) (www.anthropic.com)

For an adjacent life-sciences consultancy, the practical focus is the configured workflow. IntuitionLabs' own service description connects enterprise information to identity, permissions, retrieval, citations, evaluation, and accountable operation. That implementation perspective informs the discussion here; it does not place the consultancy among frontier-model suppliers or certify an evaluator's independence. (intuitionlabs.ai)

The analysis is current to **October 13, 2026**. The important public guidance remains differentiated: the United States Food and Drug Administration (FDA) AI credibility document is still a **January 2025 draft**, while the National Institute of Standards and Technology (NIST) TEVV-Athlon document is an **initial public draft** whose **October 6, 2026** comment deadline has passed. A closed comment period does not make a draft final.

(www.fda.gov) (www.nist.gov)

What Embedded Evaluation Adds to AI Assurance

Embedded AI evaluation means an evaluator works close to a developer's internal activities with continuing access, rather than examining only a finished model. It is an access and operating arrangement. **Independent AI assurance** is a broader claim about how evidence is produced, challenged, and communicated.

Continuous presence can improve visibility; independence still depends on contractual rights, scope control, conflicts, and reporting.

The technical distinction starts with **black-box access**, where an assessor submits inputs and observes outputs. Original research on foundation-model audits contrasts this with access to development information and model internals, which can support more targeted evaluations. Neither access pattern is automatically sufficient for every assurance question. A buyer should specify what the evaluator must conclude before deciding which access is necessary. ([arxiv.org](#)) ([arxiv.org](#))

Table 1 compares evaluation operating models. It is a proposed procurement taxonomy, informed by the UK's recommendation to combine assurance techniques and by research on access depth. It is not a ranking of providers. ([www.gov.uk](#)) ([arxiv.org](#))

Model	Timing and access	Useful assurance question	Independence and confidentiality tradeoff
External snapshot or audit	Defined assessment window; selected artifacts and query access.	Does this version satisfy the specified release criteria?	Clear engagement boundary, but conclusions depend on the evidence provided.
Recurring independent review	Scheduled reviews and change-triggered access; accumulated findings.	Have controls and performance remained adequate through changes?	Requires stable evidence rights, follow-up funding, and protected escalation.
Embedded evaluator	Continuing access to development activities and relevant internal decisions.	What does development evidence reveal beyond published outputs?	Richer visibility requires stronger access segregation, conflict controls, and reporting terms.

The distinction is consequential for pharma procurement. A periodic **red team**, meaning adversarial testing designed to expose weaknesses under specified conditions, can supply one part of the evidence. It should not substitute for review of routine scientific accuracy, record provenance, permissions, or operational controls. The UK assurance guidance explicitly treats techniques as complementary. ([www.gov.uk](#))

The announced Accenture team is to work alongside Anthropic's internal teams and safety partners. That placement identifies a developer-level engagement. It does not describe an evaluator sitting inside every pharmaceutical customer's retrieval system, examining that sponsor's production records, or signing off its intended use. Those are separate claims that require separate evidence and contractual scope. ([newsroom.accenture.com](#))

Certification also needs careful interpretation. ISO/IEC 42001 concerns establishing and improving an organizational artificial intelligence management system; ISO distinguishes that focus from the details of particular applications. An organizational assurance credential can inform supplier qualification without answering whether a specific medical-writing assistant produces accurate, traceable outputs. ([www.iso.org](#)) ([www.iso.org](#))

The proposed buying rule is therefore to combine a bounded release assessment with continuing review when changes can affect consequential decisions. Embedded access becomes valuable when the evaluator needs to examine developer processes that ordinary supplier documentation cannot explain. It should earn that access by identifying the exact claim and evidence gap, rather than by requesting unrestricted access as a default.

The Three Assurance Layers in Pharma

This report proposes **three assurance layers**: the foundation model, the sponsor-configured application, and the regulated context of use. Each has a different accountable owner and evidence package. NIST's Artificial Intelligence Risk Management Framework (AI RMF) calls for testing under conditions similar to deployment; that makes the distinction between general capability evidence and local system evidence practically important. (airc.nist.gov)

Foundation model and developer controls

The **model supplier** owns model development, release decisions, and the supplier's claims about capabilities and safeguards. Relevant evidence includes the model identifier, system card, evaluation methods, known limitations, access conditions, and findings applicable to the release under consideration. Anthropic's system-card index describes capability, safety-evaluation, and responsible-deployment documentation. A system card is useful supplier evidence, with a scope that should be read explicitly. (www.anthropic.com)

An evaluator examining this layer should state whether it tested the released model, an earlier checkpoint, or a restricted interface. Findings should identify the relevant deployment conditions and any evidence that the sponsor cannot inspect. A proposed procurement deliverable is an applicability statement linking supplier findings to the sponsor's chosen model and interface, without implying that model-level tests cover local configuration.

Configured application or agent

The **sponsor's system owner** is accountable for the implemented application: instructions, retrieval sources, tools, access permissions, memory, user interface, and operating procedures. [Anthropic's life-sciences offering](#) describes connectors, Agent Skills, and domain-specific support. These application components create an evaluation surface beyond the foundation model. (www.anthropic.com)

Model Context Protocol (MCP) connects large language model (LLM) applications with external data sources and tools. A connection can change what data an agent reads and what actions it can initiate; the protocol specification itself says it cannot enforce its security principles at protocol level. Testing must therefore include the host and downstream controls, not only generated text. (modelcontextprotocol.io)

The proposed application evidence package should include:

- **Configuration identity:** Model, system instructions, retrieval configuration, tool policies, and version history.
- **Scientific traceability:** Source passages, document versions, citation behavior, and unresolved evidence conflicts.
- **Permission evidence:** Authorized and unauthorized user paths, denied actions, and approval boundaries.
- **Workflow records:** Tool calls, resulting actions, human interventions, and relevant audit events.

This package aligns with OWASP's recommendation to preserve the tested agent version, provider, tool policy, and retrieval configuration. OWASP is a security guidance organization; its recommendations inform engineering assurance rather than granting regulatory approval. (cheatsheetseries.owasp.org)

Regulated context of use

The **sponsor's accountable process owner**, supported by quality and regulatory affairs, must connect the application to the decision it supports. FDA's draft scope concerns AI-generated information or data used in regulatory decisions about drug safety, effectiveness, or quality. That scope should not be generalized to every internal Claude workflow. (www.fda.gov)

FDA's proposed credibility framework links evidence needs to model influence on the decision and the consequence of an incorrect decision. The European Medicines Agency (EMA) separately assigns responsibility for fit-for-purpose algorithms, models, datasets, and pipelines to the relevant sponsor, applicant, holder, or manufacturer. Supplier assurance is an input into that responsibility. (www.fda.gov) (www.ema.europa.eu)

The practical boundary is explicit: a literature assistant used for orientation and an application producing evidence for a submission may share a model but require different credibility arguments. The reviewer should document the question, intended user, permitted decisions, required human review, and consequences before choosing tests. FDA describes context of use in relation to a particular question of interest. (www.fda.gov)

Evidence Access Without Exposing Confidential R&D

Access sufficiency and **data minimization** should be designed together. The US National Telecommunications and Information Administration (NTIA) identifies documentation, data, code, and models as possible evaluator inputs, subject to intellectual-property, privacy, and security protections. This supports a negotiated access plan, rather than a binary choice between a public system card and unrestricted access. (www.ntia.gov)

Table 2 is a proposed evidence-access matrix. Its disclosure limits are contract-design recommendations; they are not representations of the private Anthropic and Accenture agreement. NTIA's functional-environment discussion explains why selected technical artifacts can be necessary to evaluate a deployed system. (www.ntia.gov)

Evidence and owner	Sensitivity and evaluator access	Frequency	Disclosure limit
Development artifacts; model vendor	Proprietary methods; vendor-controlled review of relevant training and release decisions.	Release and material development changes.	Sponsor receives scope, conclusions, and exclusions; raw artifacts remain segregated.
System cards and harnesses; vendor and evaluator	Public reports plus restricted test logic; reproducible methods where permitted.	New model or revised methodology.	Share methods and results; redact protected implementation details with reasons.

Evidence and owner	Sensitivity and evaluator access	Frequency	Disclosure limit
Prompts and retrieval sources; sponsor system owner	Scientific intellectual property and potentially personal data; controlled workspace or representative test corpus.	Configuration and corpus changes.	Limit access to the approved use; prohibit onward disclosure and unrelated reuse.
Tool traces and incident logs; sponsor security and operations	Credentials, internal records, and user activity; redacted events or supervised access.	Ongoing monitoring and escalations.	Exclude secrets; separate aggregate findings from identifiable records.
Change history and findings; sponsor quality and evaluator	Controlled records; read access to approved versions and remediation evidence.	Every relevant change and closure review.	Preserve provenance, review decisions, and a record of withheld evidence.

The matrix makes an important limitation visible: a sponsor may receive an assurance summary without receiving a vendor's development artifacts. That can be acceptable if the report identifies what was examined and why the conclusions apply. It becomes insufficient when a material conclusion depends on evidence neither the sponsor nor an authorized reviewer can interrogate. The original audit-access research explains why development information can support targeted evaluation. (arxiv.org)

Confidentiality controls should cover people, environments, and outputs. METR, an independent evaluation research organization, describes project-specific siloing and access controls for model transcripts. These are concrete examples of evaluator-side segregation, not proof that every engagement uses the same controls. (metr.org) (metr.org)

For personal data, the European Commission explains that pseudonymized or deidentified records remain personal data when reidentification is possible. The European Data Protection Board (EDPB) also treats model anonymity as requiring consideration of identification and extraction risks. A sponsor should have its privacy and legal functions determine the applicable processing roles and lawful arrangements before transferring data. (commission.europa.eu) (www.edpb.europa.eu)

The proposed access contract should specify:

- **Purpose limitation:** Named evaluation tasks, permitted datasets, and prohibited secondary uses.
- **Personnel restrictions:** Approved staff, subcontractors, project segregation, and confidentiality obligations.
- **Environment controls:** Sponsor-controlled analysis where feasible, least privilege, and credential handling.
- **Retention and deletion:** Record classes, exceptions, deletion evidence, and preservation of required sponsor records.
- **Disclosure review:** Defined review period, justified redactions, and preservation of the original finding.
- **Evidence gaps:** A register of denied or partial access and its effect on confidence.

The Commission's controller-processor guidance supports confidentiality commitments for authorized personnel. The broader proposed access schedule should also cover intellectual property and scientific confidentiality, even where no personal data is present. (commission.europa.eu)

Independence, Reporting, and Remediation Ownership

Independence is a disclosed operating property, not a synonym for an external company name. Anthropic states it will fund Accenture's work directly. That fact identifies the funder; it does not settle the evaluator's scope autonomy, findings ownership, publication rights, or protection against commercial conflicts. Those questions require disclosed terms. (www.anthropic.com)

Research on **Audit Cards** proposes reporting auditor identity, scope, methodology, resource access, process integrity, and review mechanisms. The proposal is useful because an assessment result is easier to interpret when its production conditions are visible. It is a research reporting approach, not an established procurement certification. (arxiv.org) (arxiv.org)

The report proposes an independence questionnaire covering:

- **Funding independence:** Who pays, who can terminate, and whether fees depend on a favorable conclusion.
- **Scope independence:** Who selects tests and whether the assessor may investigate unexpected findings.
- **Access sufficiency:** Whether reviewers can see relevant artifacts and document denied access.
- **Publication and disclosure rights:** Who owns findings, who can edit them, and who receives unredacted conclusions.
- **Conflict controls:** Whether the evaluator also implements the system and how separate review is protected.
- **Remediation follow-up:** Whether the assessor can retest corrections and report unresolved issues.

A related public example illustrates why contractual detail matters. METR's **September 22, 2026** Claude Opus 5.5 assessment discloses an unpaid assessment agreement and a transparency provision for specified facts. METR drafted the initial summary, Anthropic had an opportunity to review and edit it, and METR signed off on the final text. The disclosed agreement includes a review and redaction process. Its transparency provision permits METR to disclose, without Anthropic's consent, the existence and general terms of that process; whether Anthropic exercised redaction rights; whether findings depended significantly on redacted information; the level of access provided; and the applicable confidentiality framework. (metr.org) This is a separate assessment disclosure, not evidence that the Accenture partnership has implemented a common embedded-evaluation standard. (metr.org) (metr.org)

Reporting should distinguish **technical findings**, **access limitations**, **management responses**, and **residual-risk acceptance**. The sponsor should receive the evaluator's conclusion alongside any vendor response, with a traceable record of redactions. The Audit Cards research supports pairing assessment claims with their justifications; the proposed reporting package extends that principle into procurement deliverables. (arxiv.org)

Accountability should be assigned through a **RACI**, meaning Responsible, Accountable, Consulted, and Informed. The following is a proposed allocation, to be adapted to the sponsor's quality system:

- **AI vendor:** Responsible for vendor artifacts and model-level corrections; accountable for its own release claims.

- **Evaluator:** Responsible and accountable for assessment methods, findings, exclusions, and retest conclusions.
- **Sponsor quality:** Accountable for the sponsor's applicable validation process and controlled assurance records.
- **Business process owner:** Accountable for permitted workflow use and acceptance of operational residual risk.
- **Platform owner:** Responsible for configuration, access controls, deployment, and application remediation.
- **Security:** Responsible for threat assessment and monitoring design; consulted on data and tool boundaries.
- **Privacy:** Consulted on data minimization, processing arrangements, retention, and assessor access.
- **Regulatory affairs:** Consulted on submission relevance, credibility evidence, and agency engagement.

The signatory for each decision must be named; a committee label alone leaves the approval ambiguous. NIST AI RMF includes regular assessment involving independent assessors or experts separate from frontline development. This supports review separation without mandating a particular consultancy or embedded model. (airc.nist.gov)

Remediation ownership should follow the source of the problem: supplier changes go to the vendor, retrieval and permissions to the sponsor's system owner, and workflow restrictions to the process owner. MITRE's SAFE-AI framework links assessment reports to remediation and reassessment of corrected controls. Closing a finding should therefore require evidence of correction and a relevant retest, not only a promise to fix it. (atlas.mitre.org)

Continuous Evaluation and Procurement Decisions

Continuous evaluation should mean maintained assurance across relevant changes, with observable triggers and accountable responses. It does not require every test to run constantly. OWASP recommends structured security testing before deployment and after material changes to prompts, tools, memory, retrieval, policies, or providers. (cheatsheetseries.owasp.org)

Table 3 translates that guidance into a proposed change-control schedule. Test scope should follow the affected assurance claim, while the approved configuration remains identifiable. (cheatsheetseries.owasp.org)

Change trigger	Evaluation response	Release or renewal evidence
Model release or routing change	Recheck applicable supplier findings and sponsor task regressions.	Model/interface identity; comparative results; updated limitations.
System prompt or memory change	Test instruction hierarchy, persistence, scientific citation behavior, and prior regressions.	Versioned prompt; memory policy; reviewed output differences.
Retrieval source or document update	Test provenance, effective-date handling, conflicting sources, and retrieval relevance.	Corpus manifest; document lineage; representative answer evidence.
MCP connector or tool change	Test data exposure, authorization, action boundaries, and tool-call traces.	Connector scope; permission tests; approved data flows.

Change trigger	Evaluation response	Release or renewal evidence
Permissions or safeguards change	Test least privilege, denied paths, high-impact approvals, and escalation.	Access matrix; control results; monitoring responsibilities.
Intended-use or policy change	Reassess decision consequence, users, human review, and required credibility evidence.	Revised context of use; accountable acceptance; regulatory consultation where relevant.

The table separates **configuration changes** from **intended-use changes**. A new use can require a different assurance argument even when the model version is unchanged. FDA's draft makes the question and context of use central; OWASP's guidance addresses material application changes. These are complementary reasons to revisit an approval. (www.fda.gov) (cheatsheetseries.owasp.org)

Connector reviews deserve particular attention. MCP's cited specification calls for explicit user consent before exposing user data to servers. OWASP recommends authorization in downstream systems and human approval for high-impact actions. A sponsor should test whether its actual implementation enforces those boundaries, including when an agent has technically valid credentials. (modelcontextprotocol.io) (genai.owasp.org) (genai.owasp.org)

Monitoring should cover task quality, source availability, permissions, unusual tool activity, human overrides, and the relevance of supplier findings. EMA's reflection paper recommends performance monitoring for drift or degradation and defined acceptable thresholds. The sponsor should determine thresholds from its use and consequences, rather than importing a generic benchmark score. (www.ema.europa.eu)

The procurement record should answer:

- **Claim purchased:** Which layer, model/interface, workflow, and decisions are within scope?
- **Evidence rights:** Which artifacts can the evaluator and sponsor review, under what restrictions?
- **Independence terms:** Who funds, sets scope, controls reporting, and verifies remediation?
- **Acceptance basis:** Which criteria must pass, and who can accept residual risk?
- **Renewal triggers:** Which changes, monitoring findings, or access restrictions reopen approval?
- **Operating commitments:** Who supplies ongoing records, performs retests, and maintains the evidence package?

A proposed decision record permits **go**, **conditional approval**, or **no-go**. Conditional approval should identify restricted uses, outstanding evidence, an owner, and a closure condition. No-go is appropriate when the buyer cannot establish the evidence or controls necessary for the intended decision. These are buyer-defined statuses, not labels conferred by the partnership announcement. NIST's ongoing-monitoring practices provide a risk-management basis for keeping decisions reviewable. (airc.nist.gov)

Data Analysis and Evidence

The relevant quantitative evidence concerns **scope and measurement**, not a league table of evaluator quality. Accenture's announcement reports expected investment of at least **\$1 billion per company over five years**. It does not disclose unit pricing, evaluator staffing, completed assessment counts, or a customer validation

success rate. Procurement teams should avoid turning the investment statement into a performance metric. (newsroom.accenture.com)

NIST's TEVV-Athlon comment period opened **August 7, 2026** for **60 days**, with comments due **October 6, 2026**. (www.nist.gov) (www.nist.gov) The initial public draft organizes work into **four stages**: Articulate & Organize, Define & Construct, Apply & Measure, and Synthesize & Interrogate. (nvlpubs.nist.gov) These establish a planning and evidence structure; a passed deadline is not a finalized standard.

FDA's draft proposes a **seven-step** credibility process: define the question, define context of use, assess model risk, develop a credibility plan, execute it, document results and deviations, and determine adequacy for that use. (www.fda.gov) Its risk reasoning incorporates model influence and decision consequence. Neither this framework nor TEVV-Athlon requires an embedded evaluator. They can inform the sponsor's assurance design within their respective scope and status.

Historical benchmark research also shows why measurement breadth matters. Stanford's **2022** Holistic Evaluation of Language Models (**HELM**) evaluated **30 models from 12 providers** and measured **seven metrics across 16 core scenarios**. (crfm.stanford.edu) (crfm.stanford.edu) Those are research-design counts, not current Claude scores or pharma acceptance criteria. The practical lesson is to report the tested scenarios and measurement dimensions, instead of presenting a single aggregate score as complete assurance.

Coverage calculation (Hypothetical Example)

The proposed **evaluation coverage** calculation is: evaluated applicable risks divided by total applicable risks, multiplied by 100. In an invented planning example, the sponsor identifies **24 applicable risks** and completes valid tests for **18**. Coverage is $18 / 24 \times 100 = 75\%$. These are author-created inputs and arithmetic, not observed industry data or an externally validated cutoff.

Coverage should be accompanied by the risk register, applicability rationale, tested configurations, exclusions, and evidence links. It says how much of the selected risk inventory was evaluated. It does not say that evaluated risks were acceptable, that untested risks are minor, or that the system is validated. NIST's deployment-context testing principle supports aligning the inventory with actual use. (airc.nist.gov)

Independence scoring (Hypothetical Example)

A proposed disclosed-answer score uses **1** for a documented satisfactory answer, **0** for a documented unsatisfactory answer, and **unknown** for undisclosed evidence. An invented contract review records: funding independence **0**; scope independence **1**; access sufficiency **1**; publication rights **unknown**; conflict controls **1**; remediation follow-up **unknown**. The report is **3 satisfactory, 1 unsatisfactory, 2 unknown**, with **3 / 4** satisfactory among disclosed inputs and **4 / 6** disclosure completeness.

No external pass mark is claimed. Unknowns stay visible, and each input retains its rationale and evidence. A sponsor can designate a particular right as essential regardless of the aggregate. This is a proposed procurement aid, informed by Audit Cards' reporting emphasis, not a comparative score assigned to Anthropic, Accenture, or another named organization. (arxiv.org)

Implications and Future Directions

Embedded evaluation may improve **developer transparency**, but pharma buyers need a separate plan for **sponsor evidence access**. The relevant question is whether a finding can support the sponsor's configured use, given the evaluator's actual access, methodology, exclusions, and reporting rights. NTIA's discussion of evaluating a model in its functional environment is directly relevant to this applicability test. (www.ntia.gov)

Data arrangements also vary by service and program. Anthropic's **Life Sciences Verification Program** announcement specifies **30-day retention** for program traffic and says retained data is compartmentalized and excluded from model training. These are program-specific statements; they should not be generalized to every Claude deployment or evaluator engagement. The announcement also states that, during the beta, LSVP is unavailable to organizations enabled for a Business Associate Agreement (BAA). (www.anthropic.com) Before including protected health information (PHI) in a proposed workflow, buyers should separately confirm a supported deployment and compliant data-handling arrangement with their privacy and legal teams. (www.anthropic.com) (www.anthropic.com)

Separately, Anthropic's **Enterprise Frontier Safeguards** announcement describes retained monitoring data in customer-controlled cloud infrastructure and a phased rollout targeting broader availability later in fall **2026**. Buyers should confirm availability and applicable deployment terms instead of assuming an announced target is an operational entitlement. This offers a potential confidentiality design pattern, rather than proof that all assurance data can remain inside every sponsor environment. (www.anthropic.com) (www.anthropic.com)

Application evidence collection may have different access implications. Claude's Enterprise documentation describes audit logs and a Compliance application programming interface (API) with access to activity records, chat histories, and file content. Procurement should map the actual licensed deployment's features to assessor needs and privacy restrictions. An available export mechanism is only useful when its scope, identity controls, and record handling fit the approved evaluation. (support.claude.com) (support.claude.com)

The next useful development would be **comparable disclosure of evaluator terms**: evidence classes, scope discretion, reporting rights, conflict safeguards, and follow-up obligations. Buyers can request this disclosure now without treating any single arrangement as the industry standard. The emerging research reporting approaches provide a basis for asking consistent questions, while organizational management standards address another part of governance. (arxiv.org) (www.iso.org)

For life-sciences implementation teams, assurance should connect to scale decisions. IntuitionLabs' published service description emphasizes measuring workflow penetration, recovered time, quality, risk signals, reliability, and support burden before expanding investment. That is an adjacent implementation perspective: sponsor evidence remains the basis for deciding where an application is useful and adequately controlled. (intuitionlabs.ai)

Frequently Asked Questions (FAQs)

Does the Anthropic and Accenture partnership validate a pharma application?

The announcements describe an embedded team evaluating models alongside Anthropic's teams. They do not present a validation package for a particular sponsor's configured workflow. Public system cards supply vendor-level documentation; buyers should request local application and context-of-use evidence separately. (www.anthropic.com)

Is independent AI assurance the same as an external audit?

An external audit describes an engagement boundary; independence describes the assessor's operating conditions. Recurring reviews and embedded evaluators can also be assessed for independence. The relevant evidence includes scope, access, methods, process integrity, and reporting mechanisms. (arxiv.org)

Does FDA require embedded evaluators for Claude?

The cited FDA document remains draft and addresses AI evidence supporting specified drug and biological-product regulatory decisions. It proposes credibility assessment for context of use, not a requirement to procure an embedded evaluator for every Claude workflow. (www.fda.gov) (www.fda.gov)

Can human review replace application testing?

Human review is part of the workflow to evaluate. Joint FDA, Health Canada, and UK Medicines and Healthcare products Regulatory Agency good machine learning practice principles emphasize human-AI team performance for medical devices. That is a useful analogy, with a device-specific scope, rather than a universal pharmaceutical requirement. Test reviewers' actual ability to detect relevant errors. (www.gov.uk)

Should every evaluator receive confidential sponsor data?

Access should match the assessment question. Controlled environments, restricted personnel, minimized records, and disclosure limits can reduce exposure while preserving relevant evidence. NTIA explicitly frames evaluator access alongside intellectual-property, privacy, and security protections. (www.ntia.gov)

Conclusion

The assurance decision is to buy **a defined evidence capability**, with a clear scope and accountable owners. The Anthropic and Accenture announcement makes embedded evaluation concrete, but the sponsor still needs an assurance argument for its configured application and intended decisions. A close relationship with the developer can provide useful visibility; it does not transfer the sponsor's responsibilities or settle the evaluator's independence terms.

The proposed framework begins with the question being answered. Model-level evidence establishes what the supplier and evaluator examined. Application evidence explains how instructions, retrieval, tools, permissions, and people affect the actual workflow. Context-of-use evidence connects those results to the decision and its consequences. Keeping these layers explicit prevents a broad supplier assessment from being interpreted beyond its documented scope.

Procurement can make the design reviewable through an evidence-access schedule, a disclosed independence questionnaire, named decision owners, and change-triggered reassessment. Confidentiality should be addressed through the same design: each access right needs a purpose, boundary, authorized recipient, retention rule, and disclosure condition. Incomplete access belongs in the confidence statement, where a decision maker can see it.

The appropriate outcome can be approval, restricted approval, or deferral pending essential evidence. Coverage percentages and independence scores should expose their inputs and unknowns, while severity and residual-risk acceptance remain separate decisions. The resulting assurance record is useful because it states what was tested, under which conditions, and when approval must be reconsidered. Supplier accountability remains with the supplier; Anthropic's announcement states that model safety remains its responsibility. (www.anthropic.com)

Source links

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

1. <https://www.anthropic.com/system-cards>
2. <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>
3. <https://www.ema.europa.eu/system/files/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle-en.pdf>
4. <https://newsroom.accenture.com/news/2026/accenture-and-anthropic-partner-to-build-team-of-embedded-evaluators-at-anthropic>
5. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>
6. <https://www.nist.gov/artificial-intelligence/ai-research/tevv-athlon-framework-evaluating-ai-systems>
7. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.200-2.ipd.pdf>
8. <https://owasp.org/about>
9. <https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/requisites-for-ai-accountability-areas-of-significant-commenter-agreement/facilitate-appropriate-access-to-ai-systems-for-evaluation>
10. https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html
11. <https://metr.org/blog/2026-02-17-how-we-protect-confidential-information/>
12. <https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance>
13. <https://www.anthropic.com/news/accenture-embedded-evaluation>
14. <https://intuitionlabs.ai>
15. <https://arxiv.org/abs/2401.14446>
16. <https://www.iso.org/standard/42001>
17. <https://intuitionlabs.ai/articles/claude-healthcare-life-sciences-ai-capabilities-2026>
18. <https://www.anthropic.com/research/claude-for-life-sciences>
19. <https://modelcontextprotocol.io/specification/2025-06-18>
20. <https://www.fda.gov/media/184830/download>

21. <https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions>
22. https://commission.europa.eu/law/law-topic/data-protection/information-business-and-organisations/application-gdpr_en
23. https://www.edpb.europa.eu/news/edpb-opinion-on-ai-models-gdpr-principles-support-responsible-ai_en
24. <https://arxiv.org/abs/2504.13839>
25. <https://metr.org/blog/2026-09-22-claude-opus-5-5/>
26. https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf
27. <https://genai.owasp.org/llmrisk/llm062025-excessive-agency/>
28. <https://crfm.stanford.edu/2022/11/17/helm.html>
29. <https://intuitionlabs.ai/articles/anthropic-life-sciences-verification-program-guide>
30. <https://www.anthropic.com/news/life-sciences-verification-program>
31. <https://www.anthropic.com/news/enterprise-frontier-safeguards>
32. <https://support.claude.com/en/articles/9797531-what-is-the-enterprise-plan>
33. <https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles>

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.