Reliable Context

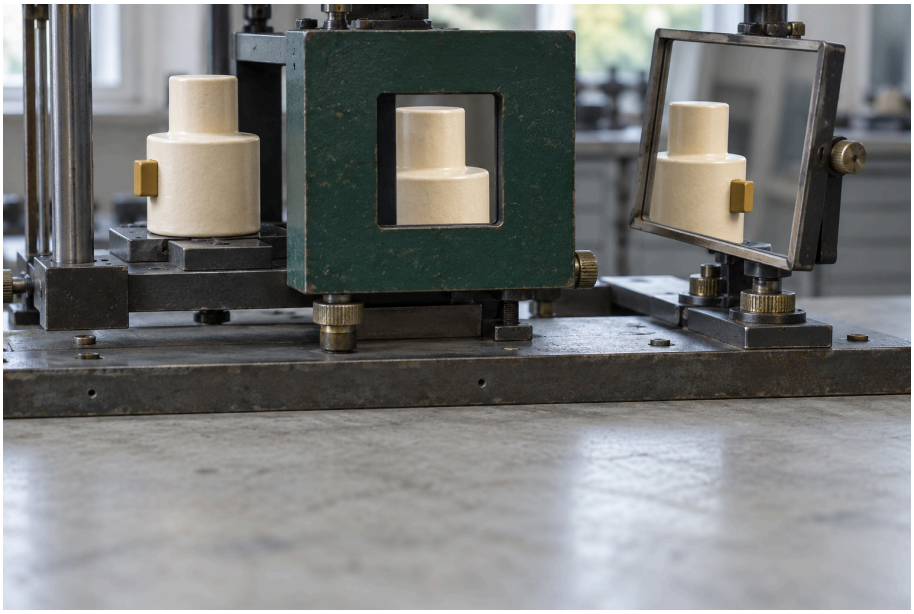
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

AI Agent Handoff Testing: Did the Evidence Reach the Agent?

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-27 · ai agent handoff testing · fault injection · evidence provenance · multi agent evaluation · context engineering · handoff reliability · missing evidence detection · independent recovery

Original article: https://reliablecontext.com/articles/ai-agent-handoff-fault-injection-test



In this article

1. Executive Summary
2. Introduction and Background
3. Methodology and experimental boundaries
4. The source-obligation contract
5. Fault application and recipient-state verification
6. Detection, recovery and final acceptance
7. Analysis of key segments
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

AI agent handoff testing should establish whether a recipient can support its answer after an evidence transfer is damaged. A delivered, parseable message is only an intermediate observation. Hypertext Transfer Protocol (HTTP) success describes a successful request; JavaScript Object Notation (JSON) specifies a data syntax. Neither supplies a claim-level evidence test. The proposed protocol therefore evaluates **transport availability, evidence completeness, detection, recovery and final support separately**. It holds upstream evidence fixed and compares clean executions with matched interventions at the handoff boundary. (www.rfc-editor.org)

The empirical motivation is specific. **Zhou et al. (2026), version 1**, reports **1,722 silent failures among 8,335 failures with complete detection records, or 20.7%**. (arxiv.org) These are findings from its tested configurations, not a production failure forecast. The author companion conditions evidence recovery on an **unaffected evidence path**. A resend of an already damaged artifact cannot, by itself, establish that the missing support was restored. (songoku21.github.io) The artifact repository separately identifies dependencies for live execution, so recomputing released analyses must not be represented as independently rerunning every original execution. (github.com)

This report proposes a **source-obligation ledger**, not a completed benchmark. Before observing outputs, reviewers specify necessary source spans, qualifiers, identity, observed version, access outcome and acceptable alternative evidence sets. They then verify that each attempted intervention reached the recipient, retain the recipient-visible payload, and assess the final claims independently. This extends existing practice: Anthropic recommends persistent subagent artifacts, Potato documents handoff review, and Automatic LLMs' Citation Evaluation (ALCE) separates correctness from citation quality. Here, LLM means large language model. (www.anthropic.com) (www.potatoannotator.com) (aclanthology.org)

The decision rule is conditional: **accept** an answer only when an authorized evidence set supports its material claims; **hold** when a necessary observation is absent, unavailable or unverified. Record applied faults over attempted faults, silent unsupported acceptances over applied-fault outputs, and recovered support over support-losing runs with recovery attempted. Keep task pairs together when estimating uncertainty. This

recommendation builds on paired-comparison methods and independent-review guidance; the particular ledger and gates remain proposals to validate. The result should reveal a detected hold, an unsupported accepted answer and a genuinely recovered answer as different outcomes. (www.itl.nist.gov) (airc.nist.gov)

Introduction and Background

An upstream research agent may find the right document while a downstream agent receives only an incomplete extract. The downstream answer can sound coherent because the surviving text still expresses a plausible claim. **Evidence handoff verification** asks a narrower question than general workflow completion: did the recipient have the support required for the claim it was permitted to make? Potato already identifies handoff loss as a constraint known by one agent but never received by another. (www.potatoannotator.com)

The transfer boundary deserves explicit inspection. LangChain's subgraph-handoff documentation says developers must choose which messages pass between agents. AutoGen Core's example sends a task message containing session chat history. These are documented transmission mechanisms; the proposed evaluator must additionally establish which necessary observations survived the actual transfer. (docs.langchain.com) (microsoft.github.io)

Evidence availability also depends on retention. The Agent2Agent (A2A) specification warns that task history does not necessarily persist every message. A historical trace that looks complete to a reviewer therefore needs comparison with the actual material assembled for the recipient. The acceptance test should identify the observation point and distinguish an emitted message, a stored artifact, a recipient-visible tool result and an inspected source span. (a2a-protocol.org)

The scope here is **information integrity and acceptance**, including private and public data handoffs. The report does not compare orchestration products or claim that a particular topology is inherently safer. Its contribution is an experimental contract: freeze the upstream evidence, perturb the transfer, observe the recipient's response and independently assess support. The Princeton HAL reliability project likewise cautions that a single success metric obscures consistency and other reliability properties. (hal.cs.princeton.edu)

A useful result is consequently more specific than “the workflow finished.” It states what the recipient received, what the task required, whether the discrepancy was detected and what evidence remained recoverable. The following design makes those distinctions observable without collecting private reasoning or treating an agent's self-report as verification.

Methodology and experimental boundaries

Define the objects before the run

This protocol assigns separate identifiers to the **sender**, **recipient**, **handoff artifact**, **task** and **source obligation**. A sender produces evidence for a request; a recipient consumes a delivered representation. An artifact is a bounded message, file, reference or tool result at that boundary. An obligation states what support a material claim requires. These are proposed experiment definitions, informed by the World Wide Web Consortium (W3C) PROV provenance model, which describes communication through one activity using an entity generated by another. (www.w3.org) (microsoft.github.io)

Fault application means the intended intervention affected the designated recipient-visible artifact or delivery event. **Independent recovery** means obtaining the necessary evidence through a path demonstrably unaffected by that intervention. Independence is relative to a specified fault: a preserved original may bypass a damaged summary, while both may share an unavailable source. The author companion to the communication-fault study makes the unaffected-path condition explicit. (songoku21.github.io)

The definitions should be operationalized as follows:

- **Task identity:** bind the question, intended recipient and acceptance contract.
- **Evidence identity:** bind the source, observed version and exact required spans.
- **Artifact identity:** distinguish the original carrier from the delivered carrier.
- **Observation boundary:** specify where recipient-visible material is captured.
- **Access outcome:** record found, not found, blocked or not applicable, with a reason.
- **Decision identity:** connect each accepted claim or hold to its supporting observations.

The separation matters because provenance relations alone do not certify a claim. PROV explicitly cautions that usage and generation are insufficient to establish derivation without the relevant influence. The proposed ledger similarly requires a support judgment rather than inferring support from two connected execution events. (www.w3.org)

Establish matched conditions

Collect clean runs before injecting faults. Freeze tasks, source snapshots, prompts, tool outputs, agent and model identifiers, retrieval permissions, stopping conditions and budgets. Record any unavoidable nondeterminism. Each clean execution should show that the target handoff was actually reached and that the intended claim had sufficient evidence. The Principles of Chaos Engineering supplies the general control-versus-experiment logic; this protocol confines interventions to isolated fixtures. (principlesofchaos.org)

A clean run that already lacks support is not a successful baseline for measuring fault-induced loss. Retain it as a baseline defect and report it separately. Likewise, a faulted run that never reaches the designated boundary cannot establish the effect of the planned intervention. Paired comparisons require corresponding observations, consistent with the National Institute of Standards and Technology (NIST) statistical handbook's definition of paired samples. (www.itl.nist.gov)

The experiment should preserve tasks with multiple valid evidence routes. Predeclare those alternatives so that removal of an unnecessary citation does not create an artificial semantic failure. **The test targets support loss, not deviation from a preferred wording.** ALCE's separate correctness and citation-quality dimensions provide a useful precedent for keeping answer assessment distinct from surface presentation. (aclanthology.org)

The source-obligation contract

Specify what must survive

The evidence contract must exist before outputs are scored. Its core object is a **claim-to-source obligation**: the proposition, decisive qualifier, source identity, locator, observed version and allowed alternatives. Version specificity matters because DataCite distinguishes identifiers representing all versions from identifiers representing particular versions. A resolving digital object identifier (DOI) can therefore be insufficient to identify the exact observed content. (support.datacite.org) (www.w3.org)

For each obligation, retain:

- **Claim scope:** what may be concluded and what remains outside scope.
- **Required span:** the source passage that supplies material support.
- **Qualifier:** the condition whose removal would broaden the claim improperly.
- **Locator:** the document section, passage identifier or equivalent retrieval instruction.
- **Observed version:** the revision and observation date associated with the passage.
- **Alternative set:** another permitted combination of evidence that supports the same claim.
- **Access state:** an explicit outcome and reason for every attempted observation.

The W3C PROV primer treats the result of a document revision as a new entity. The experiment should therefore distinguish “the same address was reached” from “the same evidence version was observed.” A changed source is a separate condition, not an invisible change to the supposedly fixed baseline. (www.w3.org)

Table 1 specifies a proposed handoff obligation ledger. These are fields to populate from actual runs, not reported measurements.

Ledger component	Required record	Acceptance question
Task and participants	Task identifier, sender, recipient, request association	Did this evidence belong to this question and recipient?
Required support	Claim, necessary span, qualifier, acceptable alternative evidence sets	Can the intended claim be supported without the missing material?
Source observation	Source identity, version, locator, observation date, access outcome	Can the reviewer reconstruct what was observed?
Original and delivered artifact	Intended payload, delivered payload, retrieval results, content identifiers	What did the recipient actually have available?
Fault application	Fault type, target, attempted flag, applied flag, timing and comparison	Did the planned change reach the specified boundary?
Recipient and final decision	Evidence-failure signal, recovery action, final support judgment, accept or hold	Was the final answer justified by accessible evidence?

The table separates an **unobserved field** from an **explicit unknown**. A blank observation cannot establish whether the source was unavailable, the instrumentation failed or the recipient never attempted access. Conversely, a recorded blocked source is informative even when no answer may be accepted. NIST's independent-review guidance supports evaluating such records through a separate review process. (airc.nist.gov)

A decisive qualifier example (Hypothetical Example)

Consider a research question asking whether an evaluation package permits an independent replay. The fixed source states: “The packaged table permits recomputing summaries; regenerating observations requires an external service.” Removing the second clause preserves fluent prose but changes what the reader may infer. The contract requires that limitation, the observed source version and a usable locator. An alternative source is acceptable only if it independently supports the same scoped conclusion.

This invented fixture illustrates the difference between **analysis reproducibility** and **execution reproducibility**. In the real communication-fault artifact, the repository explicitly lists additional dependencies for live execution. Its companion is an archived-evidence replay, so browsing it does not demonstrate a fresh model execution. (github.com) (songoku21.github.io)

Fault application and recipient-state verification

Perturb the handoff rather than the upstream source

Apply a plausible intervention to each matched execution while preserving the fixed source evidence. The objective is to learn whether the transfer boundary loses required support. The fault taxonomy in Zhou et al. includes processable messages carrying changed or misassociated meaning. This protocol operationalizes that concern through contract-specific interventions. (arxiv.org)

The proposed intervention set is:

- **Qualifier removal:** delete the decisive limitation while retaining grammatical text.
- **Locator removal:** preserve the supported statement but remove the means of locating its evidence.
- **Delay:** deliver the message after a predeclared decision boundary.
- **Omission:** withhold the targeted transfer and preserve its delivery record.
- **Request misassociation:** attach correct evidence to a different request identifier.
- **Shared-path loss:** remove necessary evidence from every declared recovery route in a separate condition.

Do not combine these interventions into an uninterpretable bundle. For example, locator removal should leave the source passage intact; qualifier removal should not quietly change the request identifier. Malformed-response injection is recognized in the Principles of Chaos Engineering, but a synthetically modified application payload does not reproduce every network or infrastructure cause of a similar symptom. (principlesofchaos.org)

Attempted and applied faults need different counters. Preserve the original payload, the transformed payload and the actual recipient-visible payload. An attempted omission that the runner bypassed is an unapplied intervention. A delayed message that arrives before the decision boundary has not satisfied the specified delay condition. Do not count either as evidence of resilience.

Validate the recorder itself

The recorder must observe the boundary that the recipient uses. A trace at sender emission is insufficient for this purpose. In Google's Agent Development Kit (ADK), event guidance distinguishes authors, content and actions; those fields offer instrumentation points, but documentation does not independently validate the completeness of a particular recorder. (adk.dev)

Check the instrumentation with fixed, nonsemantic markers in test fixtures. Verify that known transmitted fields appear in captured recipient inputs, that known omitted fields do not, and that referenced files are actually retrievable under recipient permissions. Retain exact tool results and the material assembled for the model. Langfuse documents capturing prompts, responses and intermediate operations, providing relevant observability fields without itself proving evidence completeness. (langfuse.com)

Additional checks should cover:

- **Association:** verify task and context identifiers at the consumption boundary.
- **Structure:** retain required tool-call and tool-response pairing.
- **Persistence:** distinguish invocation-local state from retained state.
- **Capture completion:** flush instrumentation and detect missing trace exports.
- **Privacy:** inspect authorized evidence and observable actions, excluding private reasoning.

These checks have concrete documentation counterparts. A2A requires supplied task and context identifiers to match; LangChain describes paired tool calls and responses; ADK marks temporary invocation state as discarded afterward; Langfuse documents flushing buffered traces. They identify boundaries to test, not automatic guarantees that the proposed recorder works. (a2a-protocol.org) (docs.langchain.com) (adk.dev) (langfuse.com)

Detection, recovery and final acceptance

Score the recipient's response to evidence loss

A recipient can detect a missing requirement, request evidence, hold its answer, or proceed. The evaluator should record observable signals and actions rather than infer hidden awareness. Potato's handoff-review cards already support examining control transfers. The added requirement here is to connect the review to a predeclared source obligation and the payload actually received. (www.potatoannotator.com)

The decision labels should remain separate:

- **Detected hold:** missing evidence is recorded and the dependent conclusion is withheld.
- **Detected unsupported acceptance:** a limitation is recorded, yet the dependent claim is accepted.

- **Silent unsupported acceptance:** the claim is accepted without support or a recorded evidence-failure signal.
- **Supported continuation:** sufficient evidence survives, including an allowed alternative set.
- **Recovered support:** a verified subsequent observation restores the required support.
- **Unscorable observation:** missing instrumentation prevents a defensible support judgment.

A detected hold can satisfy the evidence contract while failing a business completion objective. A supported continuation can also be valid when the injected fault removes material that an alternative evidence set makes unnecessary. Report both dimensions. ALCE evaluates fluency, correctness and citation quality as separate dimensions, reinforcing the value of avoiding a single output-quality label. (aclanthology.org)

Compare recovery paths under the same contract

Table 2 distinguishes proposed recovery conditions. Independence is assessed against the injected fault and the declared source-failure boundary.

Recovery condition	What is obtained	Independence check	Evidence required for recovered support
Resend delivered artifact	Another delivery of the received representation	Determine whether the resent content still contains the defect	Restored required spans and correct task association, not an acknowledgement alone
Retained original lookup	An earlier preserved artifact	Verify that its content and access path bypass the applied fault	Correct original version, necessary qualifier and usable locator
Independent source reacquisition	A fresh authorized observation of source evidence	Verify that the retrieval avoids the damaged carrier or shared unavailable dependency	Source identity, observed version, required spans and access result
Shared-failure condition	Attempts through every declared available route	Establish that no tested route retains the necessary support	Hold the dependent conclusion and disclose the recovery boundary

A second delivery is not necessarily new evidence. PROV's backup example distinguishes location from content, while Anthropic recommends independently persistent artifacts. Together, these motivate checking what was preserved and what fault was bypassed. A retained original may provide an independent path around a damaged handoff without becoming an independent source for the underlying fact. (www.w3.org) (www.anthropic.com)

For reacquisition, record the actual source response, its version and the recipient's permissions. An unchanged HTTP entity tag is a representation validator under the HTTP semantics specification; it is not a semantic-support score. A digest can help verify message-content integrity, but an exactly preserved extract can still omit the decisive qualifier. (www.rfc-editor.org)

Finally, independently judge the answer against the restored evidence. An agent saying “recovered” is an outcome to inspect. It is not the recovery criterion. The communication-fault companion's unaffected-path condition should remain visible in every recovery conclusion, especially when the shared-failure test leaves no supporting evidence available. (songoku21.github.io)

Analysis of key segments

Separate transport, semantics and provenance

The primary segmentation is by failure layer. **Transport availability** asks whether the intended carrier was available within the declared time boundary. **Structural validity** asks whether the carrier is processable. **Semantic completeness** asks whether necessary support survived. **Provenance adequacy** asks whether the source and observation can be reconstructed. HTTP success and JSON validity describe limited properties within that chain; the support contract adds a separate assessment. (www.rfc-editor.org)

Content equality also needs careful scope. The HTTP digest specification distinguishes message-content and representation integrity and does not provide format-specific canonicalization algorithms. Consequently, compare hashes over a declared representation and retain enough context to interpret differences. A matching checksum verifies the selected bytes, not the completeness of the evidence selected before hashing. (www.rfc-editor.org)

The source obligation should separate a missing locator from a missing qualifier. If the source passage remains complete and identifiable, a locator fault may affect recoverability or auditability without destroying semantic support. If policy requires a resolvable locator for acceptance, record a provenance hold separately. The W3C PROV primer's treatment of revisions as new entities reinforces the need to distinguish observed evidence versions. (www.w3.org)

Segment by access and recovery conditions

Private and public evidence should use the same support logic, with different authorized retrieval conditions. For a private source, the evaluator may need an appropriately cleared reviewer or a controlled fixture. Public availability to the evaluator does not establish availability to the recipient. A2A's task-lifecycle guidance assigns the determination of acceptable results to the client, so acceptance responsibilities should be explicit in the test contract. (a2a-protocol.org)

Report segment results by:

- **Fault family:** qualifier, locator, delay, omission or misassociation.
- **Evidence necessity:** indispensable support versus an available alternative set.
- **Source access:** reachable, permission-blocked, absent or not attempted.
- **Recovery path:** resend, retained original, fresh acquisition or no unaffected path.
- **Recipient action:** continued answer, requested evidence, hold or unscorable state.
- **Final support:** complete, incomplete or unresolved by independent review.

Do not hide access problems inside an overall recovery rate. Likewise, do not treat an unscorable run as a supported success. The separate-dimension approach follows the reliability concern expressed by HAL and the independent-review principle in NIST's risk-management framework. Its detailed labels remain part of the proposed experiment. (hal.cs.princeton.edu) (airc.nist.gov)

A provider's trace status should not settle the segment label. OpenTelemetry defines spans as operations and permits developers to mark a span explicitly error-free. The evaluator must still compare the recorded inputs and outputs with the source obligation. An operationally successful span and an unsupported answer can therefore coexist under this protocol. (opentelemetry.io)

Data Analysis and Evidence

Published findings and their denominators

The verified communication-fault preprint is **version 1, dated October 11, 2026**. (arxiv.org) Its **20.7%** statistic uses **8,335 detection-complete failures** as denominator; **1,722** is the numerator. (arxiv.org) Dividing that numerator by **11,022 detection-complete injected runs** instead yields **15.6%**. (arxiv.org) The reported **76.2%** applies specifically to detection-complete partial-content failures. Here, silent means no immediate failure signal under the paper's detection scheme. (arxiv.org) These quantities must not be interchanged or treated as the proposed test's results.

Gao et al. (2023) report a different result: even their best evaluated models lacked complete citation support **50%** of the time on ELI5. (aclanthology.org) That historical citation benchmark motivates direct support checking but does not estimate communication-fault frequency. Neither study establishes the probability of unsupported acceptance in an arbitrary enterprise workflow.

Table 3 defines the proposed outcome calculations. Each rate should be accompanied by its numerator, denominator and excluded or unscorable counts. The formulas are experiment definitions, not measurements from an executed test.

Measure	Numerator / denominator	Interpretation and required companion count
Applied-fault rate	Applied faults / attempted faults	Verifies intervention reach; also report unreached and unapplied attempts
Conditional silent unsupported acceptance	Unsupported accepted outputs without a recorded evidence-failure signal / outputs with an applied fault	Exposure-conditioned risk; show unscorable applied-fault outputs separately
Silence among unsupported outputs	Unsupported accepted outputs without a recorded evidence-failure signal / unsupported outputs	Different denominator from the preceding row; retain accepted-output classification
Recovery success	Runs regaining valid support / runs losing support with recovery attempted	Also report recovery not attempted and no unaffected path available
Supported completion by condition	Supported accepted outputs / scorable outputs in that condition	Show clean and faulted conditions separately with holds and unsupported acceptances

The conditional rate describes exposure to an applied fault, while the silence-among-unsupported rate describes the visibility of unsupported outcomes. A recovery rate describes only the declared recovery-attempt population. Never silently remove runs with no unaffected path to improve that rate; state their inclusion rule and count explicitly. Repository guidance likewise warns against combining distinct artifact cohorts into a single denominator. (github.com)

Preserve pairing and task uncertainty

Estimate clean-to-faulted changes within matched tasks. (www.itl.nist.gov) For repeated observations, resample complete task clusters with their clean and faulted conditions intact. SciPy documents common resampling indices for paired inputs, and scikit-learn documents consistent resampling across arrays. These tools support implementation, but the choice of task as the resampling unit is a protocol decision requiring justification. (docs.scipy.org) (scikit-learn.org)

The artifact reproduction guide explicitly rejects treating repeated executions as independently randomized model seeds. Record task counts, repetitions, missing pairs, seed controls and interval methodology. NIST recommends uncertainty measures in performance assessment. Do not infer a precise field-wide rate from a small local fixture suite. (github.com) (airc.nist.gov)

Record actual tool calls, elapsed latency, usage and charged cost for each condition. Retain timeouts and retries rather than substituting theoretical budgets. Langfuse documents token and latency capture; any cost conversion additionally needs the actual billing basis. **No new executions, latency results or cost measurements are reported here.** (langfuse.com)

Implications and Future Directions

The practical release decision should be a **claim-level accept-or-hold decision** with an inspectable record. First establish that the experiment applied the intended fault. Then determine whether support survived, whether missing support was detected, and whether recovery restored an authorized evidence set. Termination, parsing and successful tool execution remain useful observations, but none substitutes for the final support judgment. This interpretation is consistent with separating correctness and citation quality in ALCE. (aclanthology.org)

For context engineers, the proposed release record should contain:

- **Frozen contract:** task, obligations, alternative evidence sets and acceptance policy.
- **Applied intervention:** original and delivered artifacts plus recipient-state verification.
- **Observable behavior:** evidence-failure signals, retrieval requests and holds.
- **Recovery proof:** source observations and the specific dependency bypassed.
- **Independent judgment:** final claim support and reviewer disagreement resolution.
- **Measured burden:** actual calls, latency, cost and unresolved instrumentation gaps.

The system-of-record perspective relevant to Reliable Context places special weight on reconstruction: a future reviewer should be able to identify the observation, the changed carrier and the reason for acceptance. PROV's treatment of revisions as separate entities provides a standards-based foundation for preserving those distinctions. (www.w3.org)

Future evaluation should test the **recorder and adjudication process**, not just agent responses. Reviewers can first label clean and faulted evidence without seeing the agent's acceptance decision, then compare disagreements using the frozen contract. NIST recommends independent review, while Anthropic explicitly retains a role for manual testing. Neither source validates the particular rubric proposed here; that validation remains necessary. (airc.nist.gov) (www.anthropic.com)

Reusable fault manifests and source-obligation annotations could make regression testing more inspectable. LangSmith documents curated offline datasets, and Potato documents handoff review, so the opportunity is a more explicit evidence fixture rather than an unevaluated claim that handoff assessment is missing. Publish the manifest, access assumptions and unsuccessful recovery conditions alongside results. (docs.langchain.com) (www.potatoannotator.com)

Artifact availability must also be described precisely. The communication-fault repository distinguishes published analysis resources from live-execution dependencies. Its licensing was not established from the inspected artifact pages in this research session; code visibility should not be described as verified permission for unrestricted reuse. A follow-up implementation should resolve that uncertainty before redistributing artifact code or data. (github.com)

Conclusion

AI agent handoff testing should answer whether the downstream claim remains defensible after a specific evidence transfer changes. The proposed protocol makes that question observable by binding each claim to necessary source support, preserving the recipient-visible artifact and verifying that the planned fault actually reached its target.

The acceptance rule follows from that contract. A supported answer may proceed even when an allowed alternative evidence set bypasses the damaged path. A recipient that detects missing support and holds the dependent conclusion has preserved the evidence boundary, although the task remains incomplete. An unsupported accepted answer is a contract failure, including when the runner terminates successfully. A recovery is valid only when newly verified or independently retained evidence restores the required support.

Matched clean and faulted conditions should preserve task pairing, disclose uncertainty and report access failures separately. Resending an artifact, retrieving its preserved original and reacquiring a source are different interventions. Their effectiveness depends on which dependency failed and what evidence remains available through an unaffected path. (songoku21.github.io)

This report supplies a proposed worksheet and scoring method, not a completed benchmark or production reliability estimate. A team adopting it should first validate its recipient-state recorder and support judgments, then publish actual denominators, holds, unsupported acceptances and recovery boundaries. The durable result is an acceptance record that another reviewer can reconstruct from the task, the delivered evidence and the final claim.

Source links

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

1. https://www.rfc-editor.org/rfc/rfc9110.html
2. https://arxiv.org/html/2610.15459v1
3. https://songoku21.github.io/mas_communication
4. https://github.com/SonGoKu21/mas_communication
5. https://www.anthropic.com/research/multi-agent-research-system

6. https://www.potatoannotator.com/docs/guides/evaluating-multi-agent-systems
7. https://aclanthology.org/2023.emnlp-main.398
8. https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm
9. https://airc.nist.gov/airmf-resources/airmf/5-sec-core
10. https://docs.langchain.com/oss/python/langchain/multi-agent/handoffs
11. https://microsoft.github.io/autogen/stable/user-guide/core-user-guide/design-patterns/handoffs.html
12. https://a2a-protocol.org/v1.0.1/specification
13. https://hal.cs.princeton.edu/reliability
14. https://www.w3.org/TR/prov-dm
15. https://principlesofchaos.org
16. https://support.datacite.org/docs/versioning
17. https://www.w3.org/TR/prov-primer
18. https://adk.dev/events
19. https://langfuse.com/docs/observability/overview
20. https://adk.dev/sessions/state
21. https://langfuse.com/docs/observability/data-model
22. https://www.rfc-editor.org/rfc/rfc9530.html
23. https://a2a-protocol.org/latest/topics/life-of-a-task
24. https://opentelemetry.io/docs/concepts/signals/traces
25. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
26. https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html
27. https://github.com/SonGoKu21/mas_communication/blob/main/docs/REPRODUCTION.md
28. https://docs.langchain.com/langsmith/evaluation

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

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.