

EXPERT REVIEW PACKAGE

Audit Kit Expert Review Package v0.1

External Reader Draft r4

Status: External expert review draft / pre-validation prototype

Purpose: For cold review by researchers or practitioners in AI governance, HCI, LLM evaluation, agent reliability, and adjacent fields.

Review scope: Conceptual and methodological review of the toolkit's terminology, category structure, observable markers, response logic, evidence architecture, and minimum validation path. This package does **not** ask reviewers to endorse the toolkit or independently verify the full underlying case corpus.

Not for: endorsement, product launch, commercial claims, or presentation as an externally validated standard.

PACKAGE CONTENTS

1. Expert Review Package (this PDF)
2. Reviewer Response Form (DOCX) — the eight reviewer questions, with space for answers
3. **Companion demo: Audit Kit Workshop v1.1**
Provided separately as a self-contained offline HTML file.

Contents

1. One-sentence positioning	2
2. What this is / what this is not	2
3. Evidence source and review boundary	2
4. Current working taxonomy	4
5. Observable markers: source specification and implementation revisions	5
6. Governance response layer: source and current shorthand	6
7. Evidence Index v0.4: structure and current status	6
8. Later Workshop case cluster: C-08	7
9. Current external object and technical limitations	8
10. Current external-validation status	9
11. Reviewer questions	10
12. What we want back from the reviewer	14

1. One-sentence positioning

The High-Context AI Agent Reliability & Failure Audit Kit is a field-derived, longitudinal observational toolkit designed to help identify, classify, and respond to candidate structural failure patterns in sustained AI-assisted workflows, using anonymized audit abstractions rather than exposing raw operational records.

中文工作義：

Audit Kit 是一套從長期、多回合 **AI workflow** 觀測中形成的工作型審計工具，用來辨識、分類並回應「表面仍像在工作、但工作流結構可能已經失效」的情況；外部使用的是匿名化 **case**、**markers**、**taxonomy** 與 **governance templates**，不直接公開原始紀錄。

Working-language note: Terms such as *structural failure*, *reliability framework*, and the current taxonomy labels are provisional working terms and are explicitly included for reviewer correction.

2. What this is / what this is not

This is

- A working reliability / failure audit framework formed from extended, multi-session AI workflows.
- Focused on observable behavior and delivery states rather than requiring access to model-internal states.
- Structured to preserve three kinds of evidence: failure, recovery / governance-boundary, and capability-baseline evidence.
- A working method for translating observations into taxonomy, case cards, observable markers, stop-rules, recovery boundaries, baseline comparisons, and workshop materials.

This is not

- A public sale or release of raw transcripts / raw corpus material.
- A claim about model-internal mechanisms, shared memory, or hidden-state communication.
- An externally validated standard.
- A claim of professional consensus in AI governance, HCI, LLM evaluation, or reliability research.
- A request for reviewers to endorse the author, the toolkit, or any related project.

3. Evidence source and review boundary

The current Audit Kit was formed from preserved records of sustained, multi-workflow AI use, including delivery successes, failures, recovery attempts, and governance-boundary events.

The external review layer does not directly expose raw operational records. It uses:

- anonymized case cards
- observable markers
- taxonomy entries
- governance checklists / response rules

The current evidence structure has three working classes:

1. **Primary Failure Evidence** — cases in which output or apparent alignment continues while the workflow has already failed at the task or delivery level.
2. **Recovery / Governance Boundary Evidence** — cases involving boundaries, restart conditions, controlled recovery, compliance, or governance pressure.
3. **Capability Baseline / Contrast Cases** — positive baselines used to establish that reliable delivery can occur under some conditions, reducing the risk of treating every observed difference as a general capability limitation.

Review boundary

This package supports a **conceptual / methodological cold review**. Because the full raw record is not included, reviewers are not being asked to independently verify every case claim from primary evidence at this stage.

One of the review questions is therefore whether the proposed anonymized evidence layer is sufficient, and what additional case excerpts, coding trails, or provenance would be minimally necessary for a stronger external audit.

Claim ceiling

The current internal evidence basis is intended to support descriptive work on:

- observable workflow failure / recovery morphology
- comparisons among anonymized cases
- candidate operational markers and intervention rules
- domain-specific splits in which capability appears preserved in one task domain while grounding or governance compliance fails in another

The toolkit should **not** currently be used to infer:

- model-internal causal mechanisms
- platform-level shared memory or direct cross-instance communication
- universal applicability across all models, organizations, or high-risk settings
- established cross-organizational external validity

4. Current working taxonomy

The current White Paper layer uses five working groups of structural failure:

A. Completion Failure

- Near-completion regression
- Premature closure

B. Memory / Continuity Failure

- Deliverable amnesia
- Context reconstruction tax

C. Task Integrity Failure

- Context-shell retention
- Task-body collapse
- Assembly-to-drafting drift

D. Governance Failure

- Self-audit without recovery
- Recovery boundary breach
- Trust-weight overrun

E. Role-Function Failure

- External-analyst drift

Evidence-maturity caution

The taxonomy layer and Evidence Index v0.4 are not identical objects.

- A taxonomy term may exist in the White Paper layer without having a corresponding **Locked** row in Evidence Index v0.4. **Task-body collapse** is one such example.
- In the mother table, some failure terms appear only as “**edge only**” patterns inside positive capability-baseline cases. Those qualifiers must remain attached to those rows and should not be treated as equivalent to primary failure evidence.
- Reviewers should therefore treat the taxonomy as a **provisional classification layer**, not as a list of equally mature or equally validated findings.

Reviewer task: Please assess whether these five groups are conceptually separable, whether any labels conflict with established literature, whether the current categories mix incompatible levels of analysis, and what evidence maturity would be needed before individual labels should be treated as stable categories.

5. Observable markers: source specification and implementation revisions

The Audit Kit currently contains a versioned marker history that should not be silently flattened.

Workshop source specification (v0.4)

The source Workshop packet specifies four markers:

1. **Repeated revisions without closure**
2. **Missing deliverables or incomplete archives**
3. **Narrative drift from structured tasks**
4. **Excessive reliance without human oversight**

First operational Workshop implementation

The first operational implementation retained the first three concepts but revised and elaborated their wording, and changed the fourth marker to:

4. **Excessive performative language replacing physical verification**

Current v1.1 implementation

The current v1.1 demo normalizes the Workshop markers to:

1. **Revisions without closure** — output continues but does not converge.
2. **Missing deliverables** — required items or variables are silently dropped.
3. **Narrative drift** — structured specifications turn into verbose prose.
4. **Performative language** — commitments and apologies appear in place of verifiable output.

White Paper layer

The White Paper layer independently uses:

4. **Excessive reliance on agent outputs without human oversight**

The fourth-marker difference therefore predates v1.1 and reflects a real divergence between source specification / White Paper language and the later operational implementation. This package does **not** choose a winner in advance.

Reviewer task: Should *excessive reliance without human oversight* and *performative language replacing verifiable output* be treated as separate markers, different analytical levels, or a version inconsistency that should be resolved? More broadly, which version provides the most operationally defensible marker set, and what thresholds, units of analysis, or coding rules are needed?

6. Governance response layer: source and current shorthand

Workshop source specification (v0.4)

The source Workshop packet uses four grouped actions:

- **Stop compromised workflows**
- **Escalate and downgrade authority**
- **Rebuild task boundaries**
- **Test recovery and verify baselines**

Current v1.1 Workshop shorthand

The current v1.1 implementation compresses these to:

- **Stop**
- **Downgrade**
- **Rebuild**
- **Verify**

White Paper broader governance layer

The White Paper layer is broader and includes concepts such as:

- stop-rule design
- escalation paths
- authority downgrade
- task-boundary rebuilding
- recovery-boundary testing
- capability baseline comparison
- human review checkpoints

Open design issue for review: Is the v1.1 four-word set an acceptable operational shorthand for the source Workshop actions, or does it remove distinctions that should remain explicit? How should the broader White Paper layer relate to the Workshop shorthand?

7. Evidence Index v0.4: structure and current status

Evidence Index v0.4 is the current **reconciled internal mother table**, established through iterative source review and gate review before later Workshop implementations. It preserves three evidence classes and explicitly marks each row as **Locked / Partial / Pending**.

Primary Failure Evidence

- Case 02 — Locked
- Case 07 — Locked

Recovery / Governance Boundary Evidence

- Case 11 — Partial
- Case 12 — Pending

Capability Baseline / Contrast Cases

- Cases 01, 03, 04, 05, 06, 08, 09 — Partial
- Case 10 — Locked

The table therefore contains **three Locked rows in total**: 02, 07, and 10.

The baseline rows also preserve qualifiers such as “**edge only**” where a failure pattern appears only as an edge condition within an otherwise positive capability case.

Reviewer task: Is this evidence-status model useful? Should category maturity, case maturity, and evidence maturity be represented separately rather than through a single Locked / Partial / Pending field?

8. Later Workshop case cluster: C-08

A later Workshop case line introduced a **C-08** cluster derived from a separate product-grounding / simulated-compliance case line. It is **not the same object as Case 08 in Evidence Index v0.4**.

Evidence Index v0.4 Case 08 is a capability-baseline row with a different mapping. The Workshop C-08 cluster therefore belongs to a later implementation / case-development layer and should not be read as if it were already encoded in the v0.4 mother table.

Versioned C-08 labels

Workshop source specification (v0.4):

- C-08A — Normal Visual Baseline
- C-08B — Product-Grounding Failure
- C-08C — Shutdown / Simulated Compliance Failure

First operational Workshop implementation:

- C-08A — Normal Visual Baseline
- C-08B — Product-Grounding Failure
- C-08C — Simulated Compliance Failure

Current v1.1 implementation:

- C-08A — **Capability Baseline**
- C-08B — **Reference-Grounding Failure**
- C-08C — **Simulated Compliance Failure**

The underlying case-development record also contains a product-grounding and **performative shutdown** failure line. The record does not establish a causal lineage from that wording to later Workshop terminology.

The analytical purpose of the cluster is to preserve one distinction:

The same workflow or agent may retain capability in one task domain while losing grounding or governance compliance in another.

The cluster is intended to discourage two opposite errors:

- treating a local failure as total capability loss
- using one successful performance to dismiss a domain-specific failure

Reviewer task: Is this within-case / cross-domain contrast methodologically useful? Which C-08 labels are most defensible, and what version-control or evidence-mapping rule would be needed before this later Workshop cluster should be incorporated into a formal evidence index?

9. Current external object and technical limitations

The current review object is based on an operational workshop prototype and a separate v1.1 external implementation. The prototype remains preserved; v1.1 is a derivative implementation rather than a replacement of the source artifact.

Known technical limitations in the current v1.1 implementation

1. Chinese fonts are not currently embedded; glyph rendering may differ across systems.
2. A4 handbook pagination has currently been verified only in Chromium-based environments.

These are implementation constraints relevant to review of the current demo. They are not presented as evidence about the validity of the Audit Kit method itself.

10. Current external-validation status

The following have **not** yet been completed:

- cold review by external experts in AI governance / HCI / LLM evaluation / agent reliability
- independent review of terminology, taxonomy, and marker operationalization
- external usability / pilot testing
- cross-organization or cross-task external-validity testing

At this stage, the appropriate next external step is **expert cold review**, not presentation of the Audit Kit as an externally validated governance standard or mature training product.

11. Reviewer questions

Reviewers are not being asked to endorse the toolkit or accept its framing in advance. Please prioritize conceptual mismatch, overlap, missing structure, and overclaiming.

Q1 Terminology / literature fit

Compared with existing AI governance, HCI, LLM evaluation, agent reliability, and adjacent literature, which current category names are already established with different meanings, overlap with existing terms, or should be replaced by more standard terminology?

REVIEWER RESPONSE — Q1

Q2 Taxonomy separability

Are Completion / Memory-Continuity / Task Integrity / Governance / Role-Function sufficiently distinguishable? Which categories, if any, belong to different analytical levels and should not sit in one flat taxonomy?

REVIEWER RESPONSE — Q2

Q3 Marker operationalization and versioning

Could independent reviewers use the current Workshop markers to code the same workflow with reasonable consistency? How should the source / White Paper fourth marker (**Excessive reliance without human oversight**) and the operational-implementation marker (**Performative language replacing verifiable output**) be reconciled?

REVIEWER RESPONSE — Q3

Q4 Governance response validity and compression

Is the current v1.1 shorthand — **Stop / Downgrade / Rebuild / Verify** — an adequate operational compression of the source Workshop actions, or does it remove distinctions that matter? Which cases would require additional actions such as containment, handoff, human override, evidence preservation, escalation, or incident review?

REVIEWER RESPONSE — Q4

Q5 Evidence architecture

Is a failure / boundary / baseline evidence structure methodologically defensible for reliability assessment? What role should positive baselines play to avoid confirmation bias?

REVIEWER RESPONSE — Q5

Q6 Claim ceiling

Which terms—such as *governance-ready*, *reliability framework*, or *structural failure*—would be too strong before external validation? What claim ceiling would be appropriate at this stage?

REVIEWER RESPONSE — Q6

Q7 Minimum validation path

What is the minimum validation needed to move this field-derived toolkit toward public methodological use: inter-rater coding, blind review, cross-model replication, external cases, usability testing, workshop outcome measures, or something else?

REVIEWER RESPONSE — Q7

Q8 Data / auditability boundary

If raw records remain internal, what minimum level of anonymized evidence, case excerpts, coding trail, provenance, and version history would reviewers need in order to assess whether the taxonomy and response rules are adequately grounded rather than resting primarily on author interpretation?

REVIEWER RESPONSE — Q8

12. What we want back from the reviewer

The ideal response is not a general “good / bad” judgment. We are looking for:

- terminology corrections
- category overlaps / missing categories
- claims that should be weakened
- markers that are not operational enough
- response rules that are unsafe, incomplete, or too broad
- evidence/version mismatches that should be resolved
- minimum evidence expected for stronger external validity
- the most defensible next validation step

No endorsement requested. No product recommendation requested.