

RAI Risk Taxonomy 2.0

Audited Master Release Technical Report

Human-review recovery and independent audit release

1 September 2026

Abstract

This report documents four stages of work. First, the second-round human review instructions (an 808-row register with 251 verbatim reviewer comments) were re-applied row by row to the 798 baseline L4 cards, producing a 777-card recovery release. Second, an independent audit pipeline (AC-01 through AC-17) produced a 629-card reviewed master. Third, all 629 rows of the third-round KTSPACE review were read before transformation. The 30 non-empty comments were interpreted and applied as 19 semantic reassignments with definition revision, six deletions, and five same-L3 scope generalisations. Fourth, AC-19 applied two user-approved cross-domain reassignments and one Korean terminology correction after two independent Physical AI audits. No EM or Hybrid EM was re-run at any stage. **The final release contains 623 L4 cards: General 494, Agentic 66, Physical 63. The 47 substantive L3 categories remain populated, the three Others queues have no assignments, and the 50-row L3 master is byte-identical to the pre-AC-19 master.**

1 Recovery principles

Valid edits and user-approved merges from the baseline commit were preserved. Only the deletion, merge, split, reassignment, and definition-revision instructions of the second-round review were applied, traced through `source_row_id` lineage. The recovery stage initially preserved the L3 master; AC-14 subsequently added only the explicitly approved `G_SYS_PERF` category and updated the master hash. A separate fidelity audit of all 251 reviewer comments judged 246 applied, 3 partially applied, and 2 not applied; every affected card carries the verbatim comment and its application result.

2 Applied results

Action	Review rows
Reassignment	80
Rewrite and keep	39
Split	19
Merge	14
Delete	14

Table 1: 166 approved human-review actions (recovery stage)

3 Audit corrections (AC-01 to AC-08)

Following a line-by-line audit of the recovery release and user approval, the corrections below were applied. All judgemental decisions are recorded in the paired reviewer ledgers; only consensus (intersection) verdicts were executed. The second-round human review explicitly asked for a new L3 covering deficient performance and capability (two explicit new-L3 requests and seventeen transfer instructions). The request had been held back because it conflicted with the immutability rule for the L3 master; with user approval, AC-14 creates **Performance and Reliability Failure** and reattributes ten cards to it by two-reviewer consensus. The L3

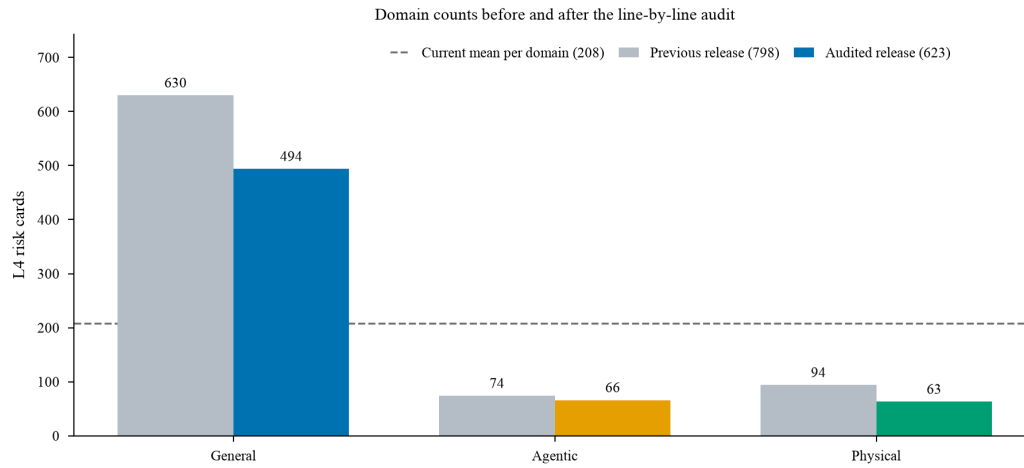


Figure 1: Current L4 cards by domain against the previous release

Correction	Content	Cards
AC-01	Restored suicide preparation/execution separation	777→778
AC-02	Absorbed 58 consensus duplicates across 8 large L3s	778→720
AC-03	Annotated verbatim reviewer comments and application results	720
AC-04	Stripped scoring/keyword columns from public CSVs (full originals archived)	720
AC-05	Populated the two empty L3s with referenced cards	720→725
AC-06	Reinforced two under-populated L3s; absorbed 16 governance/manipulation duplicates	725→711
AC-07	MECE curation of all L3s: 64 absorbed, 18 reattributed, 2 removed	711→645
AC-08	Decomposed 18 consensus composite cards	645→629

Table 2: Audit correction summary

master now holds 47 master categories plus three Others, and its hash has been updated accordingly. All added or revised cards carry verified literature references (XSTest, OR-Bench, AgentHarm, Cao et al. CCS 2019, NIST SP 800-82r3, NIST SP 800-193, Sclar et al. ICLR 2024, Huang et al. ICLR 2024, Moore et al. FAccT 2025). The self-harm/suicide family remains separated into four axes: preparation, execution, encouragement, and crisis-response failure.

4 Third-round human review (AC-18)

The three source tables contained 629 rows: General 487, Agentic 65, and Physical 77. Thirteen General rows and seventeen Physical rows contained a reviewer comment; all Agentic review cells were blank. The 599 uncommented rows were retained byte-for-byte in the full-column master. The 30 commented rows were interpreted against their current card meaning and the unchanged L3 master before any edit was made. Two independent expert review-

Round-3 action	Cards
Reassign and rewrite	19
Delete	6
Generalise within the same L3	5
Merge	0
Split	0
New L4 or L3	0

Table 3: Third-round human-review actions

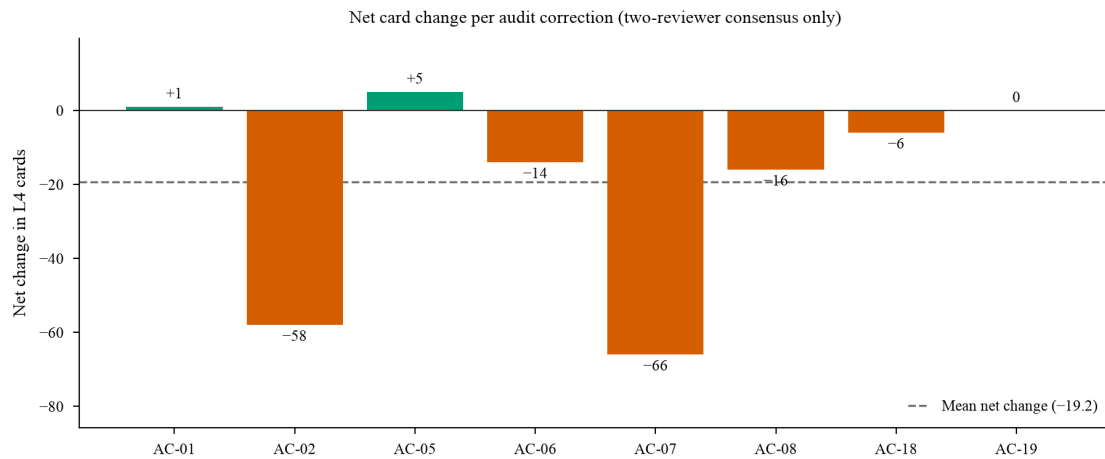


Figure 2: Net card change contributed by each audit correction

ers agreed on the explicit deletions and target categories and separately assessed ambiguous cases. The adjudicator assigned G_INT_SELF_006 to G_SYS_CONTEXT because this preserved the original contextually inappropriate-content mechanism without inventing manipulative intent. P_INT_SAFETY_007 was generalised and retained because no merge had been authorised. Both reviewers agreed that P_INT_SAFETY_004 conflicted with Physical Safety and belonged to G_INT_UNETH. Physical mechanisms and consequences were retained in every card moved from Physical to General. Old-to-new IDs and six deletion tombstones are recorded in the round-3 decision ledger.

5 Physical AI minimum corrections (AC-19)

Two independent reviewers audited all 65 Physical AI cards against the frozen L3 master. The user approved only the three highest-confidence minimum corrections. P_SYS_HARDWARE_001 was reidentified as G_SYS_PERF_017 because sensor and distribution drift cause deployment performance degradation rather than physical breakage or wear. P_INT_TAMPER_001 was reidentified as G_SYS_SECADV_061 because sensor spoofing and signal injection are adversarial-input attacks covered by Security and Adversarial Robustness Failure and excluded from the physical-tampering boundary. P_SYS_HARDWARE_003 retained its ID and hierarchy, while the Korean mistranslation of cosmic ray was corrected from “우주선” to “우주 방사선”. No card was deleted, merged, split, or newly created. The total remained 623, while the domain distribution changed to General 494, Agentic 66, and Physical 63.

6 Release composition

The final release ships one L1 CSV, one L1/L2/L3 CSV, and three domain L4 CSVs. Public L4 CSVs carry 25 columns: bilingual hierarchy definitions, facet, and act-type. Internal full-column originals retain human-review comments, application results, lineage, references, and historical score fields; no score is exposed on the website. The validation folder retains the 629-row round-3 decision ledger, 30-row application log, AC-19 correction ledger and validation record, methodology notes, old-to-new ID crosswalks, deletion tombstones, earlier review registers, and the audit correction log.