RAI Risk Taxonomy 2.0: Top-down evidence selection and staged classification of global AI risks

RAI Risk Taxonomy Project

22 July 2026 · journal edition, release state v2.17.2

Abstract

This report gives a reproducible, top-down account of how RAI Risk Taxonomy 2.0 was constructed, classified, and validated. A multi-million-record evidence space of papers, policy reports, and patents is filtered into a curated AI-risk evidence view, from which a permanent registry of canonical L4 risk cards is built and assigned to a fixed set of semantic L3 families under a strict L0–L4 hierarchy; the previously validated Physical AI taxonomy is preserved throughout as a locked gold set. Classification proceeds in stages: a strict open-set gate that releases only unanimous assignments, a suitability-ranked relaxation, and a forced-path completion, followed by independent expert review, guarded constrained-EM audits, and a registry-preserving consolidation that retires only verbatim duplicate records while keeping every identifier citable. No algorithmic or forced placement is treated as ground truth: each remains flagged for human adjudication through an explicit HOLD review state. Validation combines permutation tests, centroid-containment analysis, embedding-perturbation stability, and encoder-transfer checks. The evidence shows that the taxonomy carries statistically non-random semantic structure, that the HOLD mechanism effectively concentrates weak-fit and ambiguous cases, and that consolidation of duplicates measurably tightens the semantic geometry, while global reliability of the single-parent partition remains under review. We argue that residual HOLD concentration reflects cross-cutting risks and category overlap rather than mere misclassification, and we specify every evidence filter, classification rule, and validation limit needed for replication.

Keywords: AI risk taxonomy; evidence synthesis; open-set classification; semantic mapping; Physical AI; responsible AI.

1 Introduction

A taxonomy can appear complete while concealing uncertainty in its source corpus, unit of analysis, or placement rules. RAI Risk Taxonomy 2.0 therefore separates four quantities that must not be conflated: the size of the integrated AI evidence space, the size of the AI-risk evidence view, the number of canonical L4 risk cards, and the number of L4 cards assigned to L3. The central result is a top-down evidence-to-taxonomy chain of 3,906,767 records, 82,971 AI-risk records, 1,726 candidate L4 rows, 1,725 exact-deduplicated candidates, 1,711 level-validated L4 cards, and 1,711 L3-assigned cards.

The transitions do not all preserve the same statistical unit. The first two stages count documents or patent records. The third and fourth stages count canonical risk concepts. Consequently, the 82,971 evidence records are not a one-to-one parent table from which every L4 card was mechanically generated. They form an auditable evidence view supporting corpus coverage and risk interpretation, whereas the L4 registry integrates risk taxonomies, benchmarks, standards, and linked evidence through a source crosswalk.

The remainder of this paper is organized as an academic report. Section 2 describes the top-down construction of the evidence corpus and the canonical L4 registry. Section 3 specifies the classification methodology, including the staged hierarchy-blind procedure, the guarded constrained-EM audit, the registry-preserving consolidation pass, and the HOLD review policy. Section 4 reports release composition and reliability validation through the current release state (v2.17.2). Section 5 interprets the validation evidence and its limits, and Section 6 states the release decision and review priorities. Version-by-version engineering history, detailed review policies, exploratory analyses, and reporting conventions are collected in the appendices.

2 Evidence corpus and canonical L4 registry

2.1 Integrated AI evidence space

The top-level dataset contains 3,906,767 records from three source families: 1,749,159 papers, 125,372 policy reports, and 2,032,236 patents. Paper records originate from the local global Web of Science AI collections. Policy records combine the reviewed curated AI-policy master with the broader policy-materials collection. Patent records originate from the stage-2 AI patent master.

Records are standardized to a common schema and retained only when a non-empty title is available. Deduplication is performed within document type and analytical category. The ordered key hierarchy is patent application ID for patents, followed by DOI, source record ID, URL, and a normalized title-year-institution fallback. Cross-category overlap is intentionally retained because AI, AI infrastructure, Physical AI, and AI risk are analytical views rather than mutually exclusive partitions.

2.2 Integrated AI-risk evidence view

The AI-risk evidence view contains 82,971 records: 50,375 papers, 938 policy reports, and 31,658 patents. Its inclusion rule is source-governed rather than a single keyword query. Papers combine curated academic risk references, the clean and integrated AI-risk paper corpora, Physical AI safety benchmarks, and verified paper records from the top-200 evidence set. Policy reports combine curated risk-policy references with verified report records. Patents combine curated risk-patent references with verified patent records. Record-type filters are applied before the same title-validity and within-category deduplication rules used at Stage A.

This view represents 2.1% of the integrated AI evidence space. Because the source collections and analytical categories can overlap, this percentage is a descriptive ratio, not the estimated prevalence of risk research in all AI scholarship.

2.3 Canonical L4 registry

The unit changes from evidence record to risk concept. Source risk IDs are joined through the source crosswalk, semantically redundant cards are reviewed, and one canonical row is retained per L4 failure mode or harm mechanism. Permanent identifiers `RAI4-0001` through `RAI4-1726` are allocated independently of the risk label. The initial result is 1,726 candidate L4 rows. Automatic deduplication is restricted to pairs whose normalized label and mechanism-only definition are both exactly equal; semantic-similarity thresholds are not used. This merges `RAI4-1606` into `RAI4-1083`, retains both source references, and yields 1,725 exact-deduplicated candidates. The retired ID remains in the crosswalk. The registry includes the 182-card Physical AI taxonomy as a protected mapping set.

The ratio of L4 cards to AI-risk evidence records is 2.1%. It is reported only as a scale-change diagnostic: one document may support multiple risks, and one risk may be supported by multiple documents. It is not a document acceptance rate.

2.4 Taxonomy-level validation

The deployed Physical AI site is the authoritative source for its 24 L3 names and 182 L4 cards. Its live HTML was byte-identical to the local repository snapshot used for this check. We compare the English component of each authoritative Physical L3 name with every candidate L4 label after NFKC normalization, case folding, ampersand normalization, and removal of non-alphanumeric separators. A card is excluded only when these normalized labels are exactly equal; semantic similarity and definition similarity are not used. Fourteen L4 rows match category names and are retired as taxonomy-level leakage, including `RAI4-0553` “Accidental harm”, which maps to Physical L3 `P3.1`. Their definitions and references remain in the audit crosswalk. This yields 1,711 canonical L4 cards without changing any of the protected 182 Physical L4 mappings.

2.5 Assignment coverage overview

The operational policy assigns all 1,711 canonical L4 cards to an L3 node (100.0%). The original three-stage classifier used 50 L3 nodes and retained 76 decision-review markers after level validation. Release v2.7.0 then removed 1,129 instances of circular taxonomy prose from non-Physical definitions, performed evidence-constrained mapping review, and marked 692 cards where the available definition and evidence did not justify an approved taxonomic fit. The v2.8.0 Agentic expansion evaluates every non-Physical card while locking all 182 Physical mappings. It moves 31 cards into four new L3 nodes and clears seven review markers, leaving 685 cards (40.0%) marked HOLD. Releases v2.9.0 to v2.12.0 then audit Anthropomorphism, Goal Misalignment, and Copyrights, leaving 689 cards (40.3%) marked HOLD. Release v2.13.0 differentiates three overlapping Memory cards and places all three under source-entailment or near-duplicate scope review, yielding 692 held cards (40.4%). Release v2.14.0 then retires the four sparse Agentic families and force-migrates all 39 affected cards; 28 newly enter review, yielding 720 held cards (42.1%). The marker is internal review metadata on an already assigned card, not a separate L1, L2, or L3 category.

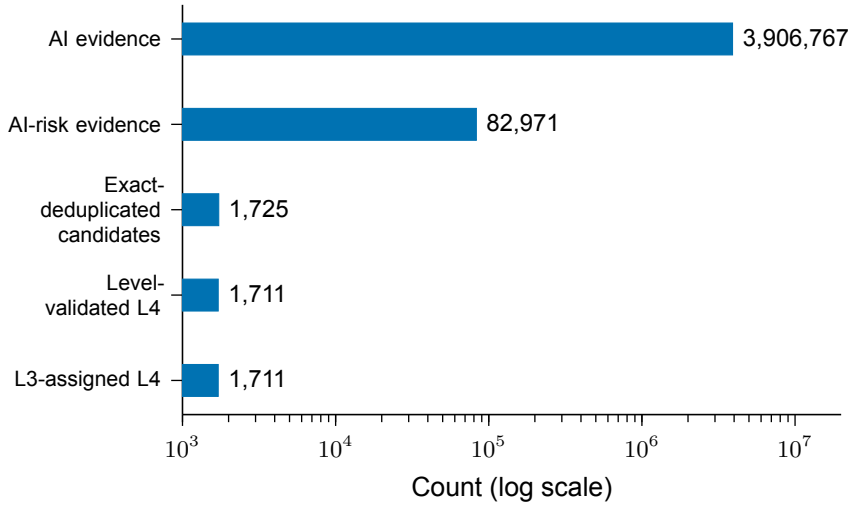

Figure 1: **Top-down evidence and taxonomy funnel.** Counts move from document/patent records to canonical risk concepts at the transition from AI-risk evidence to L4. The logarithmic axis preserves visibility across four orders of magnitude.

Table 1: Top-down stages, counts, and inclusion filters.

Stage	Count	Unit	Principal filter
AI evidence	3,906,767	Document or patent record	AI source collections; valid title; standardized schema; within-type/category deduplication.
AI-risk evidence	82,971	Document or patent record	Curated/integrated risk sources; record-type inclusion; valid title; within-type/category deduplication.
Exact-deduplicated candidates	1,725	Risk card	Exact normalized label and core-definition match only; one retired ID mapped to its canonical card.
Canonical L4	1,711	Risk card	Exact normalized-label exclusion against 24 authoritative Physical L3 names; 14 level leaks retired; all 182 Physical L4 cards preserved.
L3 assigned	1,711	Risk card	Protected Physical mapping or staged non-Physical placement; 689 decision-required cards remain assigned and marked HOLD.

3 Classification methodology

3.1 Hierarchy and protected Physical AI mapping

The taxonomy uses three canonical L2 categories (Interaction Safety, System Safety, Societal Safety) plus a HOLD review overlay, and permanent `RAI4-####` identifiers that are independent of taxonomy placement (Appendix A).

The 182 L4-to-L3 mappings inherited from the Physical AI taxonomy are treated as the locked gold set. Later stages cannot change these paths. Before the taxonomy-level exclusion, the remaining 1,543 candidates enter the hierarchy-blind classification procedure. Legacy global L0-L3 labels, source families, and source-ID prefixes remain reference metadata and are excluded from prediction features.

3.2 Stage 1: strict open-set gate

The input for each non-Physical card is the L4 label, a mechanism-only definition with hierarchy restatements removed, and the evidence title. Let s_{sem} be the top-1 semantic score, s_{comp} the top-1 composite score, m_{sem} and m_{comp} their respective top-1 margins, and v_{anchor} the number of anchor views selecting the same L3. A candidate is retained only when all seven conditions hold:

1. exactly one L3 is rule-eligible;
2. the rule winner, semantic top-1, and composite top-1 are identical;
3. $s_{sem} \geq 0.55$ and $s_{comp} \geq 0.54$;
4. $m_{sem} \geq 0.035$ and $m_{comp} \geq 0.035$;
5. $v_{anchor} \geq 2$ across three anchor views;
6. no gap sentinel, multi-mechanism flag, or Physical-boundary flag is present; and
7. two independent expert agents approve the same exact L3.

This gate yields 37 new expert-consensus proposals. Together with the 182 Physical locks, Stage 1 assigns 219 cards and withholds 1,507.

3.3 Stage 2: suitability-ranked relaxation

Stage 2 preserves all 219 protected assignments and computes

$$S_2 = 0.55q_{sem} + 0.20q_{sm} + 0.10q_{cm} + 0.10q_{vote} + 0.05q_{rule} - p_{gap}, \tag{1}$$

where

$$q_{sem} = \text{clip}\left(\frac{s_{sem} - 0.45}{0.25}, 0, 1\right), \tag{2}$$

$$q_{sm} = \text{clip}\left(\frac{m_{sem}}{0.05}, 0, 1\right), \tag{3}$$

$$q_{cm} = \text{clip}\left(\frac{m_{comp}}{0.05}, 0, 1\right), \tag{4}$$

$$q_{vote} = v_{anchor}/3, \tag{5}$$

$$q_{rule} = \mathbf{1}[\text{at least one eligible L3}], \tag{6}$$

$$p_{gap} = 0.05\,\mathbf{1}[\text{gap sentinel present}]. \tag{7}$$

S_2 is a deterministic ranking score rather than a calibrated probability. Stage 2 first preserves 55 hard holds. It then orders the remaining holds by ascending S_2, with permanent RAI4 ID as the tie breaker, and reserves the lowest 118. Thus $55 + 118 = 173$ cards remain unresolved and 1,553 cards are assigned (90.0%).

3.4 Stage 3: forced-path completion

Stage 3 preserves the 1,553 Stage 2 paths and assigns only the 173 unresolved cards to the hierarchy-blind top-1 L3 already stored in the score artifact. This produces 1,726 operational paths but does not validate taxonomy fit. Each new path retains its original hold reason, score, review priority, and `human_approved=false` marker.

The final policy marks five difficult-case families for decision review: 23 Physical-scope reviews, 17 expert rejections, 7 expert disagreements, 6 multi-mechanism cases, and 2 low-absolute-fit cases. Their forced L3 path remains populated and `decision_required=true`. The remaining 118 quota holds also retain their forced paths and review metadata.

3.5 Constrained-EM audit procedure

Later audits do not apply unconstrained nearest-centroid output. Each audit allows only the selected source L3 cards to move, treats all other cards as fixed anchors, excludes every Physical destination, and recomputes candidate centroids until no assignment changes. The composite score is

$$S(i,k) = 0.60\, c(i,k) + 0.30\, d(i,k) + 0.10\, w(i,k), \tag{8}$$

where c is current-centroid cosine similarity, d is L3-definition cosine similarity, and w is TF-IDF keyword cosine similarity. A destination must improve on the source by at least 0.020 and have keyword cosine of at least 0.015. Agentic destinations additionally require an explicit autonomous execution mechanism, and destination-specific definition guards reject matches caused by generic shared terms.

The cumulative audits move three Anthropomorphism cards in v2.9.0, five Goal Misalignment cards in v2.10.0, four Copyrights cards in v2.11.0, and eight additional Anthropomorphism cards in v2.12.0. Every such move remains `HOLD` until human adjudication. The final domain counts are 1,165 General, 364 Agentic, and 182 Physical cards; the review queue contains 689 cards.

3.6 Definition revision and guarded audit (v2.17.0)

Release v2.17.0 applies an expert editorial pass to the non-Physical registry while keeping all 182 Physical cards byte-identical. The pass rewrites 147 title and definition mismatch candidates identified by the v2.16.0 remediation audit into mechanism-only bilingual definitions, renames 15 non-descriptive labels with the original label preserved in provenance, applies 187 mechanical hygiene fixes (hyphenation artifacts, missing final punctuation, duplicated tokens) to the remaining non-Physical definitions, and attaches five externally verified governance references (NIST AI 100-1; OWASP Top 10 for LLM Applications 2025) to directly matching cards. Every textual change carries `human_review_status=pending` and retains the original text in a `definition_revision` record; no HOLD marker is cleared by this pass.

The revised card text is then re-audited with the constrained-EM procedure of Section 3.5 using the pinned local BGE-M3 snapshot `5617a9f61b028005a4858fdac845db406aefb181`. The unguarded BGE-M3 control run produced 151 candidate moves and converged after six non-zero update passes, followed by a seventh zero-move pass. This result is intentionally used as sensitivity evidence rather than as automatic ground truth: it shows that the edited definitions expose many plausible local alternatives, but the volume of candidates is too high for direct publication without human review. The published v2.17.0 placement layer therefore preserves the guarded RC policy: 22 defensible proposals are recorded, of which 14 produce actual `primary_l3_id` changes and 8 update the semantic path of cards that were already in a HOLD review path. All 22 movers carry `decision_required=true`; the review queue grows from 720 to 734. No Physical path changed.

The apparent difference between 22 proposals and 14 primary moves is therefore structural, not an accounting error. A non-HOLD card stores its operational L3 in `primary_l3_id`, so a guarded proposal changes that field. A card that is already in a HOLD path keeps `primary_l3_id` at the HOLD node and stores the proposed semantic family in `hold_semantic_path`. Those eight cards are valid reassignment proposals but do not appear as primary L3 moves in a direct card-to-card diff.

3.7 Registry-preserving consolidation and label-noise remediation (v2.17.2)

Release v2.17.2 addresses five classes of L4 label noise identified during review: overly generic names, exact or near-duplicate names, parenthetical or compound labels mixing description and scope, domain-topic-like names, and multi-mechanism labels. Every exact normalized-label cluster was adjudicated card by card against the mechanism definitions. Records whose label and mechanism definition are verbatim or same-lineage duplicates (for example, double-indexed 2021/2022 editions of the same survey, or condensed copies of a source taxonomy) are retired with a `merged_into` pointer, union of references, and alias preservation; the identifier registry is invariant, and no record is deleted. Cards with an articulable subtle difference are preserved and receive unique mechanism-bearing labels, with the distinction recorded in provenance; twelve initially merged records were revived under this preservation-first rule. Multi-mechanism cards receive a primary mechanism label with the secondary axes stored in a `secondary_mechanisms` field, consistent with the multi-label extension discussed in Section 5. All label and definition changes carry `human_review_status=pending` with the original text preserved.

3.8 HOLD review policy

`HOLD` is the public rendering of `decision_required=true`. In v2.15.0 it is a review-path L2 under General and Agentic AI, not a semantic risk family. A held card has one navigable HOLD path and retains its prior semantic L3 candidate as explicit lineage metadata; neither path is human-approved ground truth. The state is set by a hard qualitative trigger or by a pending algorithmic-review trigger. It is cleared only by a recorded human decision that reconciles the label, mechanism-only definition, evidence, and destination definition; absence of a competing score alone is not sufficient.

Quantitative criteria. Quantitative measurements control automatic eligibility and review priority, not semantic truth. The following rules are applied or reported:

1. Stage 1 automatic acceptance requires $s_{sem} \geq 0.55$, $s_{comp} \geq 0.54$, both margins at least 0.035, at least two of three anchor votes, exact agreement of the rule, semantic, and composite winners, and two matching expert-agent approvals. Failure prevents automatic clearance.

2. Stage 2 uses the deterministic suitability score S_2 defined above; the lowest-ranked reserve and every hard hold retain decision metadata after forced operational assignment.

3. A constrained-EM move requires composite improvement of at least 0.020 over the source and keyword cosine of at least 0.015. Passing these thresholds permits a proposal but never clears `HOLD` automatically.

4. For v2.12.0 audit prioritization, a top-two cosine margin below 0.020 is high priority, a margin from 0.020 to below 0.050 is medium priority, and a margin of at least 0.050 is lower priority. Released-L3 rank outside the top three or per-card agreement below 80.0% under $\sigma = 0.05$ perturbation is an additional instability signal.

5. Strong geometric support is reported only when top-1 containment is at least 90.0%, top-2 containment at least 95.0%, the matched-size permutation test has $p < 0.001$, and perturbation agreement remains at least 97.0% through $\sigma = 0.05$. These are validation thresholds, not substitutes for the qualitative tests below.

The qualitative retention criteria are listed in Appendix B.

4 Results

4.1 Evidence composition

Figure 2 compares the document-type composition of the integrated evidence space with that of the AI-risk evidence view. The risk view is markedly more paper-weighted than the full space, which is dominated

by patents. This asymmetry matters for interpretation: patent language describes mechanisms and applications rather than harms, so the curated risk view supplies most of the harm-mechanism vocabulary that the downstream card registry draws on.

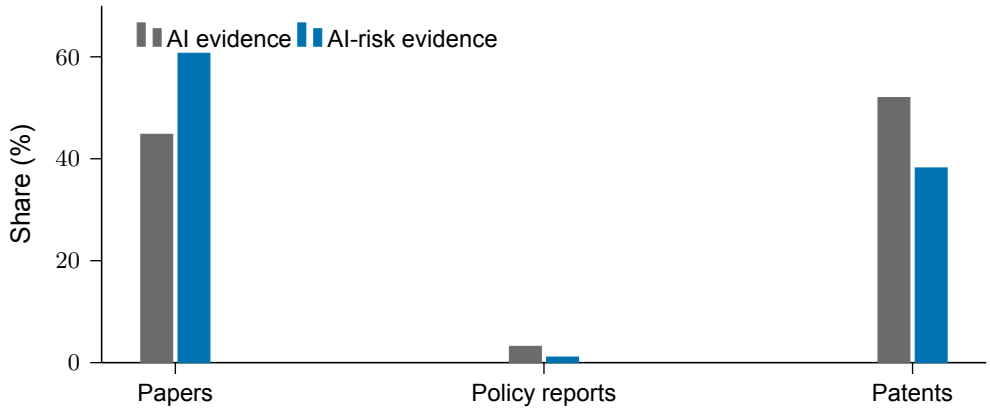

Figure 2: **Source-family composition.** Percentages compare document-type composition within the integrated AI evidence space and the AI-risk evidence view.

4.2 Hierarchy coverage

The final policy assigns all 1,711 canonical cards (100.0%); 720 assigned cards (42.1%) carry the HOLD marker. Of all cards, 1,193 are General, 336 are Agentic AI, and 182 are Physical AI. Their shares are 69.7%, 19.6%, and 10.6%, respectively. All 50 semantic L3 nodes contain at least one non-HOLD card. The two HOLD review-path L3 nodes contain 626 General and 94 Agentic cards, respectively. Anthropomorphism is the largest L3 with 234 cards (13.7%); the five largest L3 nodes jointly contain 41.6%.

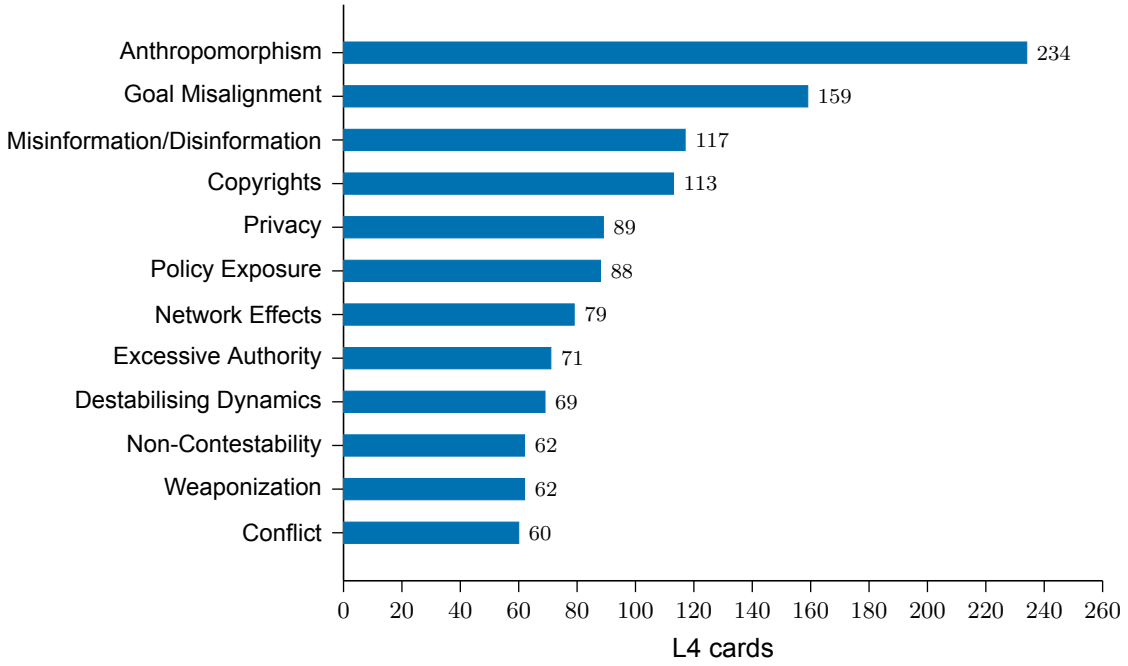

Figure 3: **Largest L3 nodes in the final policy view.** The figure shows the eleven largest L3 nodes. Counts include assigned cards marked HOLD.

After the v2.17.1 minimal Physical transfer and the v2.17.2 consolidation (Appendix A), the public release comprises 1,711 registered IDs, 1,660 active cards, 189 Physical-path cards, and 765 active HOLD records.

4.3 HOLD composition at v2.17.0

The v2.17.0 HOLD queue contains 734 cards, or 42.9% of the 1,711-card release. HOLD is not a semantic failure label; it is an operational review state assigned when the evidence, L4 mechanism, and candidate L3 definition do not jointly justify automatic release as a human-approved mapping. To make the review queue actionable, the release groups HOLD reasons into mutually exclusive operational causes using `decision_reason`, `stage2_hold_reason`, and the v2.17.0 guarded remap metadata. The machine-readable output is stored in `reports/validation/v2.17.0/hold_reason_groups/`.

Table 2: Primary operational causes of HOLD status in v2.17.0.

HOLD cause	Cards	Share of HOLD	Share of all L4
Taxonomy gap or missing L3 scope	280	38.1%	16.4%
Anthropomorphism mechanism ambiguity	173	23.6%	10.1%
Overloaded L3 or weak evidence fit	138	18.8%	8.1%
Expert disagreement or rejection	51	6.9%	3.0%
Retired sparse Agentic L3 migration	39	5.3%	2.3%
Guarded BGE-M3 remap review	26	3.5%	1.5%
Physical AI lock-scope conflict	16	2.2%	0.9%
Low-fit or multi-mechanism residual	8	1.1%	0.5%
Other residual review	3	0.4%	0.2%

The distribution shows that HOLD is dominated by three causes. First, 280 cards expose a taxonomy gap or missing L3 scope. These cards are not simply bad assignments; many point to governance, transparency, security, dependency, environmental, labor, and deliberate-disinformation themes that are only partially covered by the current 50 semantic L3 families. Second, 173 cards remain held because an Anthropomorphism assignment is not established as a direct harm mechanism. These cards frequently mention trust, identity, affective influence, or social interaction, but the definition does not prove that anthropomorphic design is the operative risk mechanism. Third, 138 cards sit in overloaded L3 families or have weak evidence fit. Their current L3 may be operationally usable, but the source evidence is too broad or the L4 mechanism too diffuse for direct release.

Table 3: Taxonomy-gap mentions inside the HOLD queue. Multi-tag cards are counted once per tag.

Taxonomy-gap theme	Mentions	Share of gap mentions
Governance, audit, and regulation	87	28.8%
Transparency and explainability	44	14.6%
Robustness, alignment, and control	34	11.3%
General AI security	31	10.3%
Dependency and manipulation	30	9.9%
Pluralistic values and stakeholders	21	7.0%
Environment, energy, and resources	20	6.6%
Labor, inequality, and power	17	5.6%
Deliberate disinformation	16	5.3%
Other taxonomy-gap tags	2	0.7%

A HOLD-only BGE-M3 remap probe was also run for the 734 held cards. It found 496 cards whose top-1 semantic L3 differs from the stored HOLD semantic path, but all 496 are weak-tier candidates under the conservative threshold: no card met the joint score, margin, and keyword-support requirements for automatic movement. This confirms that HOLD should be used to prioritize human review, not to trigger bulk remapping. The highest-priority review queue should start with taxonomy-gap cards, then Anthropomorphism ambiguity, then overloaded-L3 or weak-evidence cases.

4.4 Reliability validation procedure

Reliability is re-evaluated on all 1,711 v2.14.0 paths using dense, 1,024-dimensional BGE-M3 embeddings, seed 20260721, maximum sequence length 256, batch size 32, and L2 normalization. Card text concatenates bilingual L4 labels and definitions; L3 seed text concatenates bilingual L3 labels and definitions. The procedure reproduces four checks from the prior Technical Report. This rerun encodes the current v2.14.0 card text and all 50 active L3 seed definitions. It therefore incorporates both the v2.13.0 Memory wording revisions and the v2.14.0 forced migrations.

First, seed-initialized spherical EM alternates nearest-centroid assignment and normalized centroid recomputation:

$$z_i^{(t+1)} = \arg\max_k y_i^\top \mu_k^{(t)}, \qquad \mu_k^{(t+1)} = \frac{\sum_{i:z_i=k} y_i}{\left\| \sum_{i:z_i=k} y_i \right\|_2}. \tag{9}$$

Second, top-k containment is the proportion of cards whose published L3 is among the k nearest centroids calculated from the published partition. Third, the observed mean within-family cosine is compared with 5,000 label permutations that exactly preserve family sizes; the p-value uses the plus-one correction. Fourth, isotropic Gaussian embedding perturbation is repeated 200 times per σ, and agreement compares perturbed with unperturbed nearest-centroid assignments.

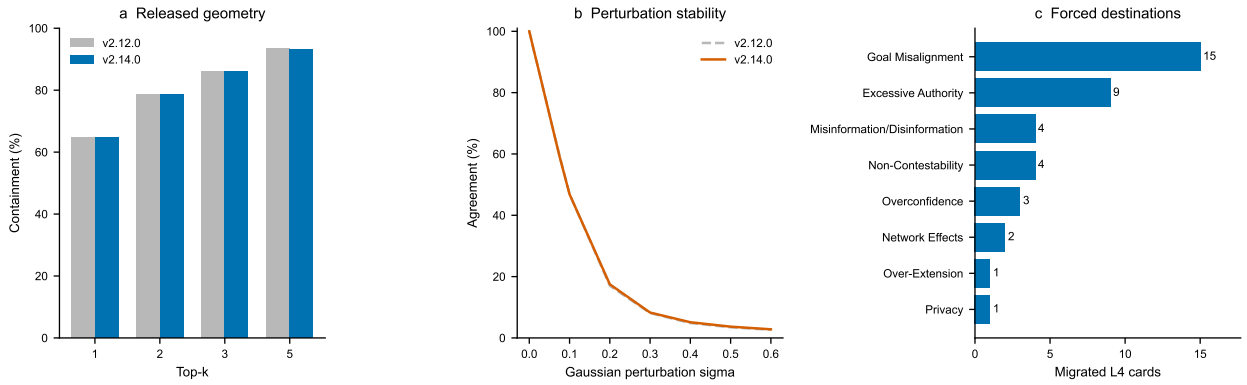

Figure 4: **Convergence and stability before HOLD separation (v2.14.0).** All 1,711 cards are assessed while the 720 HOLD cards remain embedded in their semantic L3 candidates. (a) Seed-initialized spherical EM reaches a fixed point after 15 iterations. (b) v2.14.0 top-k containment is compared with the v2.12.0 54-family baseline. (c) Perturbation stability remains materially similar after consolidation.

Table 4: Full-release reliability results before and after L3 consolidation.

Metric	v2.12.0	v2.14.0
L3 families	54	50
Cards assessed	1,711	1,711
Mean within-family cosine	0.770	0.768
Median assignment margin	0.012	0.012
Top-1 containment	64.6%	64.7%
Top-2 containment	78.7%	78.6%
Top-3 containment	86.1%	86.1%
Top-5 containment	93.3%	93.1%
Permutation p-value	0.0002	0.0002
Agreement at $\sigma = 0.01$	94.4%	94.5%
Agreement at $\sigma = 0.05$	72.8%	72.8%

The EM objective increases monotonically from 0.661 to 0.800 and reaches zero reassignments after 15 iterations. The converged unsupervised partition, however, agrees with only 21.8% of published paths (adjusted Rand index 0.120; normalized mutual information 0.402) and proposes 1,338 changes. Numerical

convergence is therefore not evidence that the published semantics are correct. Published cohesion is 0.7684, above the matched-size null mean of 0.7382 with zero exceedances in 5,000 permutations (plus-one $p = 0.0002$). The taxonomy contains non-random semantic structure, but its global partition is not robust under the tested geometry.

4.5 Reliability of the v2.17.0 audit

The final sensitivity analysis uses the canonical BGE-M3 encoder, seed 20260721, the same 50 semantic L3 families, 5,000 matched-size permutations, and the same perturbation grid used in the historical analyses (Appendix D). Table 5 reports the geometry before and after the BGE-M3 constrained audit. HOLD is reported both included and excluded because the HOLD marker is a review state, not a semantic L3 family.

Table 5: BGE-M3 reliability before and after the v2.17.0 constrained-EM audit.

Metric	All, pre	All, post	Non-HOLD, pre	Non-HOLD, post
Cards assessed	1,711	1,711	977	977
Top-1 containment	64.9%	69.5%	76.8%	79.3%
Top-2 containment	78.9%	83.5%	88.3%	91.6%
Top-3 containment	85.9%	90.4%	92.2%	94.6%
Top-5 containment	93.0%	95.0%	95.6%	97.4%
Median assignment margin	0.0125	0.0146	0.0240	0.0279
Negative-margin share	35.1%	30.5%	23.2%	20.7%
Permutation p	0.0002	0.0002	0.0002	0.0002
EM convergence iterations	25	25	12	12
EM objective at convergence	0.801	0.801	0.802	0.802
EM agreement with released paths	20.6%	22.3%	30.0%	31.7%
Agreement at $\sigma = 0.01$	94.1%	94.7%	95.6%	96.0%
Agreement at $\sigma = 0.05$	72.3%	74.0%	78.4%	80.0%

Three conclusions follow from Table 5 without restating its entries. First, the permutation test rejects random labeling under BGE-M3 in every condition, so the partition encodes real semantic structure rather than lexical accident. Second, excluding the HOLD queue improves every indicator simultaneously, which is the signature of an effective triage mechanism: the review state absorbs the cards whose definitions genuinely underdetermine a single family. Third, the constrained audit improves local geometry but cannot, by construction, validate the taxonomy globally: unconstrained spherical EM lacks the hierarchy, the protected Physical lock, and the review policy, so its disagreement with released paths measures the difference between a purely geometric partition and a normative one, not an error rate. The release conclusion therefore remains extbfSHARE WITH CAVEATS, with the guarded reassignment proposals and the HOLD queue as the priority adjudication targets.

4.6 Current release state (v2.17.2)

Release v2.17.2 incorporates the final L4 label-deduplication and redefinition pass. The ID registry is invariant at 1,711 registered L4 IDs. Of these, 1,660 records are displayed and validated as active cards, while 51 records are retained as merged provenance records with `merged_into` links. The site and report therefore state both numbers together: **1,711 registered IDs, 1,660 active cards**. This preserves citation, audit, and crosswalk stability while preventing exact or same-lineage duplicate records from being treated as independent active risks.

The BGE-M3 constrained-EM mapping algorithm was rerun on the 1,660 active cards using the unchanged 50 semantic L3 families from v2.17.1. The 189 Physical-path cards are locked. The rerun produced 144 guarded move candidates. Following the release policy, these candidates are not treated as human-approved ground truth: every reflected move remains a HOLD review item. For cards already in HOLD, only the stored semantic review path is updated; for previously non-HOLD non-Physical cards, the primary

path is moved into the appropriate General or Agentic HOLD review path and the proposed semantic L3 is stored in `hold_semantic_path`. This yields 765 active HOLD records.

Table 6: BGE-M3 reliability after the v2.17.2 L4 revision pass.

Metric	All, pre	All, post	Non-HOLD, pre	Non-HOLD, post
Cards assessed	1,660	1,660	954	954
Top-1 containment	65.5%	70.2%	77.3%	79.7%
Top-2 containment	78.8%	84.0%	89.0%	91.3%
Top-3 containment	86.3%	89.9%	92.1%	94.5%
Top-5 containment	92.9%	94.6%	96.0%	97.2%
Median assignment margin	0.0116	0.0150	0.0258	0.0276
Negative-margin share	34.5%	29.8%	22.7%	20.3%
Permutation p	0.0002	0.0002	0.0002	0.0002
EM convergence iterations	15	15	19	19
EM objective at convergence	0.797	0.797	0.800	0.800
EM agreement with released paths	23.1%	25.7%	29.7%	31.4%
Agreement at $\sigma = 0.01$	94.1%	94.5%	95.5%	96.2%
Agreement at $\sigma = 0.05$	72.0%	73.4%	78.8%	79.8%

Table 6 shows that the consolidation and guarded remap pass improves every reliability indicator on both populations. The mechanism of the gain is informative: retiring verbatim duplicates removes near-identical points that previously split family mass across identical intensions, and mechanism-bearing relabeling moves card text closer to the definitional anchor of its family. The improvement is therefore a data-quality effect, not an algorithmic recalibration, and it supports publication as a synchronized review release while every reflected move remains a HOLD review item pending human adjudication.

4.7 HOLD sensitivity at the current release state (v2.17.2)

The HOLD sensitivity analysis of Appendix D is repeated here at the current release state so that the triage claim is not supported only by a superseded release. The procedure applies the BGE-M3 active-card embeddings to three separately evaluated scopes: the full active release with Physical AI included and HOLD included, the same full active release after removing HOLD cards, and the Physical AI subset alone. Retired merged records remain preserved in the 1,711-ID registry but are excluded from reliability metrics because they are no longer independent active risk cards.

Table 7: Three-scope sensitivity analysis at v2.17.2 (BGE-M3, active cards only, 5,000 permutations, 200 perturbation repeats per σ).

Scope	Cards	EM steps	Objective	Top-1	Top-2	Top-3	Top-5	$\sigma = 0.05$
Overall, Physical included, HOLD included	1,660	15	0.797	70.2%	84.0%	89.9%	94.6%	73.4%
Overall, Physical included, HOLD excluded	895	14	0.799	81.0%	92.0%	94.7%	97.1%	80.4%
Physical AI only	189	6	0.835	90.5%	96.8%	99.5%	100.0%	87.1%

The qualitative pattern of the historical analysis replicates under the canonical BGE-M3 encoder after the v2.17.2 consolidation pass: removing the review queue improves containment, margin quality, and perturbation stability together, while non-random structure holds in both full-release conditions. The Physical AI subset remains the strongest and most stable reference population. The correct reading remains conditional: the non-HOLD subset is cleaner because the policy deliberately concentrates weak-fit and multi-mechanism cases in the queue, so the comparison validates the separation signal, not the accuracy of every individual non-HOLD assignment. Machine-readable outputs are stored in `reports/validation/v2.17.2/three_scope_sensitivity_bge_m3/`.

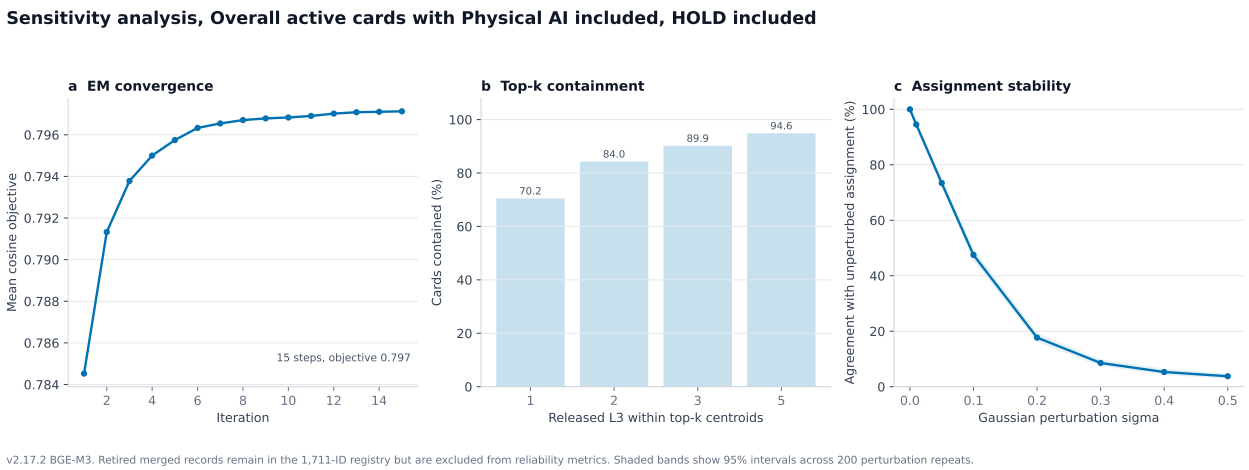

Figure 5: Sensitivity analysis for the full v2.17.2 active release with Physical AI included and HOLD included. (a) Seed-initialized spherical EM reaches a fixed point after 15 iterations. (b) Released-L3 top-k containment is evaluated across $k = 1, 2, 3, 5$. (c) Assignment stability is measured under Gaussian embedding perturbation. Shaded bands indicate 95% intervals across 200 perturbation repeats.

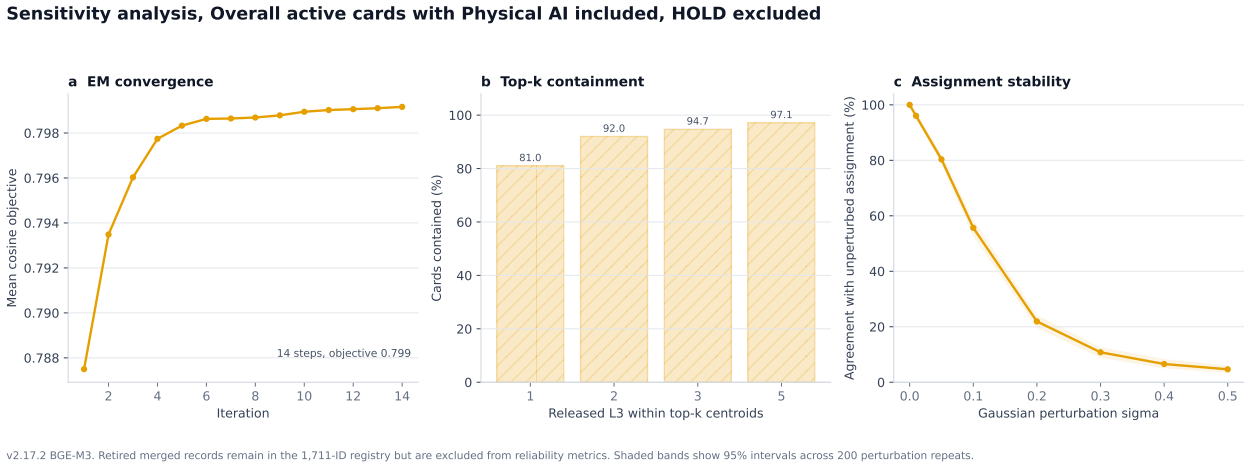

Figure 6: Sensitivity analysis for the full v2.17.2 active release with Physical AI included and HOLD excluded. Removing HOLD cards increases Top-1 containment from 70.2% to 81.0%, increases Top-5 containment from 94.6% to 97.1%, and increases assignment stability at $\sigma = 0.05$ from 73.4% to 80.4%.

4.8 Policy-subset separation

The current HOLD queue is measurably weaker than the non-HOLD subset under the same released-centroid geometry. This supports the review marker as a triage instrument, although it does not prove that every individual marker is correct.

Table 8: v2.14.0 reliability by policy subset.

Subset	Cards	Mean cosine	Median margin	Top-1	Top-3
Physical locked	182	0.817	0.038	89.0%	97.3%
Non-Physical	1,529	0.763	0.009	61.8%	84.8%
HOLD	720	0.759	0.004	56.1%	82.6%
Non-HOLD	991	0.775	0.018	70.9%	88.6%
Forced migrations	39	0.746	-0.005	46.2%	71.8%

The subset ordering in the table matches what the review policy predicts: the locked Physical set is

Figure 7: Sensitivity analysis for the Physical AI subset alone. The 189 active Physical AI cards converge in 6 EM iterations, reach 90.5% Top-1 containment and 100.0% Top-5 containment, and maintain 87.1% assignment stability at $\sigma = 0.05$.

strongest, the non-HOLD subset intermediate, and the HOLD queue and forced migrations weakest. Two interpretive cautions apply. The Physical result evidences internal alignment of the authoritative set but is not an unbiased comparison, because its paths are locks rather than predictions. The HOLD result validates the marker as a triage instrument without proving any individual marker correct, since the policy itself concentrates weak-fit cases in the queue. Successive releases changed these subset statistics only marginally, so the release verdict rests on the qualitative separation rather than on any single decimal.

5 Discussion

5.1 What the validation does and does not establish

Sparse-family consolidation: OPERATIONAL PASS WITH REVIEW. Exactly 39 cards were migrated, no Physical path or L4 identity changed, all four retired families were archived, and every migrated card remains HOLD. The weak migrated-subset geometry prohibits interpreting the destinations as approved ground truth.

Full 50-family taxonomy: NOT YET DEMONSTRATED. The permutation test establishes non-random structure, but global containment, perturbation stability, and EM-to-published agreement do not satisfy strong reliability criteria. The weakest top-1 families are Destabilising Dynamics (36.2%), Goal Misalignment (38.2%), Network Effects (41.6%), Conflict (50.0%), Copyrights (53.1%), and Inconsistency (57.1%). Anthropomorphism remains large and heterogeneous at 234 cards and 65.0% top-1 containment.

Four limitations constrain interpretation. First, cross-category overlap in the evidence space means that source counts are analytical views, not mutually exclusive samples. Second, the transition from 82,971 records to 1,711 L4 cards changes the unit of analysis. Third, embedding geometry is a diagnostic, not a substitute for expert judgment or causal evidence. Fourth, small families can obtain perfect containment with limited statistical power.

5.2 Independent specialist diagnosis

Two specialist agents independently reviewed the validation artefacts. One review concentrated on representation learning, spherical EM, and statistical estimands; the other examined ontology design, data provenance, and review policy. They were instructed to separate observations from causal inference and falsifiable hypotheses. Their agreement is an analytical review, not a substitute for the independent human gold-standard study recommended below.

The reviewers first cautioned that the three headline values estimate different properties and are not in-

terchangeable accuracy measures. The 64.6% top-1 value is an *in-sample centroid-containment rate*: every card contributes to the released-family centroid against which it is evaluated. This construction is optimistic for small families and tautological for singletons, while its micro-average is heavily influenced by large families. The 72.9% agreement at $\sigma = 0.05$ measures sensitivity to a specified, uncalibrated correlated Gaussian perturbation of cosine scores; it is not an estimate of encoder reproducibility or assignment correctness. Finally, the 20.4% EM-to-release agreement compares an unconstrained spherical clustering objective with a normative expert taxonomy. It does not imply that 79.6% of released paths are erroneous.

Table 9: Two-reviewer diagnosis of the reliability estimands.

Diagnostic	Why the value can be limited	What the value does not establish
Top-1 containment, 64.6%	Overlapping L3 intensions, heterogeneous broad families, single-centroid misspecification, and in-sample family-size effects	Independent accuracy or a 35.4% error rate
$\sigma = 0.05$ agreement, 72.9%	The perturbation scale is about four times the median margin of 0.0123 and is not calibrated to observed encoder or text variation	Real-world reproducibility, semantic validity, or correctness probability
EM-to-release agreement, 20.4%	EM has no hierarchy, protected-path, family-size, causal-mechanism, or multi-label constraints and may drift from its seed meaning	That 79.6% of published paths should be moved
Permutation $p = 0.0002$	Large N yields a narrow null around a high cosine baseline; observed cohesion exceeds the null by 0.0314	Correctness, completeness, or uniqueness of the 54 families

The principal explanation is ontological rather than purely computational. The L3 layer mixes harm content, failure mechanisms, affected rights, governance processes, and interaction dynamics in one exclusive partition. Many L4 risks are therefore multi-axial: anthropomorphic trust can co-occur with privacy disclosure, manipulation, and non-contestability. Forced primary-only placement converts legitimate semantic multiplicity into narrow or negative margins. Boundary overlap is especially plausible between Goal Misalignment and Goal & Planning; Excessive Authority, Tool Calling, and Oversight & Control; and Network Effects, Destabilising Dynamics, Conflict, and Collusion. A single prototype per family cannot represent such multimodal intensions reliably.

The observed family distribution supports this diagnosis. Family sizes range from 1 to 234, with a median of 15.5; 19 families contain at most five cards, whereas 12 contain at least 50. Family size and top-1 containment are negatively correlated ($r = -0.70$), as are family size and median margin ($r = -0.52$). Anthropomorphism contains 234 cards, of which 229 remain `HOLD`; 173 lack an established anthropomorphic mechanism. Goal Misalignment contains 144 cards and has 38.2% top-1 containment with a negative median margin. Taxonomy gaps, an unestablished Anthropomorphism mechanism, and overloaded or low-fit destinations account for 591 of 689 HOLD cards (85.8%). These concentrations are consistent with large residual families absorbing risks for which a stable destination is absent or insufficiently defined.

Definition provenance is a second structural limitation. Among 1,529 non-Physical cards, 1,510 use the `source_mechanism_only_v2.7` definition form and 1,528 cite exactly one reference. Earlier remediation removed 1,129 instances of explicit upper-taxonomy scaffold prose, but truncating that prose does not by itself demonstrate that the retained atomic definition is entailed by the cited source. The embedding audit may consequently measure consistency of editorial abstractions as well as evidence-grounded risk identity. The stronger Physical result must also be interpreted separately: its 182 paths are authoritative locks with synchronized definitions, and 16 of its 24 families contain at most five cards. Its 89.0% top-1 result supports internal alignment but is not an unbiased comparison with non-Physical assignments.

The two reviews converge on four testable hypotheses. First, independently coded primary-plus-secondary mechanisms should be more common among low-margin cards if single-label pressure is causal. Second, leave-one-out or held-out centroids should reduce the apparent advantage of singleton and small families. Third, source-entailment rewriting should improve expert agreement and held-out margins if definition provenance is causal. Fourth, paraphrase, translation, truncation, and encoder-version experiments should yield an empirical score-drift distribution against which perturbation scales can be calibrated. Until

those tests are completed, the current metrics support targeted review rather than wholesale algorithmic remapping.

Table 10: Reviewers' agreed validation priorities.

Priority	Intervention	Acceptance evidence
P0	Construct a stratified human gold set by L1, L3 size, HOLD, and boundary status; allow one primary and optional secondary L3	At least two domain experts per card, adjudication, agreement coefficient with uncertainty, and micro/macro accuracy with confidence intervals
P0	Rewrite L3 intensions as inclusion rules, exclusions, and contrastive examples; audit every sampled label-definition-reference triple for direct source entailment	At least 80% blinded agreement on boundary contrast sets; unsupported cards rewritten, split, or retired
P1	Review the 591 HOLD cards in the three dominant reason groups and resolve Anthropomorphism residuals without automatic cluster splitting	Documented migration matrix and repeated before/after reliability analysis
P1	Enforce Agentic-necessity predicates and evaluate primary plus secondary mechanisms	Generic model risks remain General; boundary-review agreement at least 0.70
P2	Replace resubstitution scoring with leave-one-out or held-out prototypes; report macro and micro estimates, confidence intervals, and empirical text perturbations	No self-inclusion; families with $n < 5$ marked insufficient; noise scale calibrated from observed variation

5.3 Cross-cutting risks under a single-parent hierarchy

Many L4 risk cards exhibit cross-cutting risk characteristics and therefore become boundary cases under the current single-parent hierarchy: anthropomorphic trust, for example, can simultaneously implicate privacy disclosure, manipulation, and non-contestability. The concentration of such cards in the HOLD queue reflects category overlap between L3 intensions rather than necessarily incorrect classification. A principled future extension is to treat these cards through multi-label classification, in which one mandatory primary family and optional secondary families are coded, or through a polyhierarchical taxonomy in which an L4 card may hold more than one L3 parent. The `secondary_mechanisms` field introduced in v2.17.2 provides the first data support for this extension while preserving single-point operational accountability through the mandatory primary parent.

5.4 Candidate L3 expansion

A conservative candidate-L3 analysis and a counterfactual reliability simulation (Appendix C) indicate that adding four General and two Agentic candidate families creates plausible movement candidates but does not improve top-k containment, margin quality, or perturbation stability. The candidates therefore remain review hypotheses; no expansion is adopted in the current release.

6 Conclusion and review priorities

The release conclusion is **SHARE WITH CAVEATS**: the taxonomy exhibits statistically non-random semantic structure, an effective HOLD triage separation, and measurable reliability gains from the v2.17.2 consolidation, but strong global assignment reliability is not yet demonstrated.

Review should begin with the v2.17.2 HOLD queue, especially the 144 guarded BGE-M3 move candidates, then the weakest global families and low-margin cards. The 765 active cards marked `HOLD` should be adjudicated using agreement among the risk label, mechanism-only definition, cited evidence, and candidate L3 definition. Each approved batch should be followed by the identical reliability analysis. A low score alone must not trigger automatic movement; human review determines whether to retain the path, expand an L3 definition, or propose a genuinely necessary new L3.

Appendix A Release history and engineering decisions

A.1 Authoritative Physical L4 content synchronization (v2.4)

Release v2.4 synchronizes the content of all 182 protected Physical cards from the local repository underlying the deployed Physical AI Risk Taxonomy site. The immutable RAI4 crosswalk and L3 paths are retained, while the bilingual label and definition, severity, probability, 3H1R attributes, evidence justifications, and 360 reference rows are replaced by the current Physical source values. Of these reference rows, 359 contain a justification and one source row retains an empty justification without imputation. Bilingual terminal parenthetical text is split into Korean and English fields to preserve the global card schema. Impact is recalculated as $I = s\,p$ from the synchronized severity s and probability p. Percentile is not displayed for any card because the cards do not constitute a common measurement population suitable for relative ranking. Existing percentile fields remain provenance data only.

A.2 Complete English–Korean localization (v2.5)

Release v2.5 requires bilingual display at every taxonomy level. The supplied L0-L3 terminology and Korean definitions are used as the controlled glossary. The 182 Physical L4 cards retain their authoritative bilingual source text. For the remaining 1,529 cards, the localization pipeline removes the repeated provenance boilerplate from each English definition, selects the first sentence that states the core failure mechanism, generates a concise Korean definition, and normalizes risk-domain terminology. Context-sensitive corrections include *poisoning* as "data or model contamination" rather than substance addiction, *washing* as false compliance presentation, and *rubber stamping* as nominal approval. Every generated row retains the English source-sentence hash, method, and `human_review_status=pending`. Thus all 1,711 cards contain non-empty English and Korean labels and definitions, while automated Korean drafts remain explicitly reviewable rather than being represented as human-approved ground truth.

A.3 Canonical L2 consolidation (v2.6)

Release v2.6 normalizes L2 to three canonical categories: *Interaction Safety*, *System Safety*, and *Societal Safety*. The former General labels *Interaction* and *System*, and the former Agentic label *System*, are renamed accordingly. Six L1-specific path nodes remain as implementation edges so that the strict L0-L1-L2-L3 tree and all L3 parent paths are preserved; each edge references one of the three canonical L2 identifiers. Thus the reported L2 category count is three, while no L3 path or L4 assignment changes. Release v2.15.0 adds a fourth canonical display category, HOLD, with General and Agentic path nodes. This fourth category is a review-state overlay rather than a semantic safety dimension.

A.4 Stable L4 identifiers and L1 display semantics

The public explorer keeps the stable `RAI4-####` identifier independent of taxonomy placement. L1 membership is communicated through an explicit text badge, a consistent icon, and a visual color: a brain and blue for General-purpose AI, a compass and green for Agentic AI, and a robot and red for Physical AI. The icon-color pair is repeated in the global navigation, filters, hierarchy panel, and L4 cards. The color is derived from the card's current L1 path at render time rather than encoded into the permanent identifier. This allows later human-approved remapping without identifier churn and avoids using color or icon as the sole carrier of categorical information.

A.5 Independent expert card review

Two reviewers independently inspected all 1,725 cards for agreement among the L4 label, mechanism definition, evidence, and assigned L3. Reviewer A proposed 30 amendments and Reviewer B proposed 14. Automatic amendment required both reviewers to propose the same label and L3 at high confidence. One card met this rule: `RAI4-0888` changed from "Privacy concerns" to "Anthropomorphic trust-induced privacy

disclosure", while retaining RAI3-G-INT-10. Its original label remains in provenance metadata. Forty-two cards had a proposal from only one reviewer or incompatible proposals; their labels and paths were not changed. Thirteen were already marked HOLD and 29 newly received the marker. The consensus amendment cleared the prior hold on RAI4-0888, giving $55 - 1 + 29 = 83$ final HOLD cards. No protected Physical AI card was changed.

The later taxonomy-level validation supersedes the proposed adjudication of RAI4-0553: "Accidental harm" is itself the authoritative Physical L3 category P3.1, not an L4 failure mode. It is therefore excluded rather than renamed or remapped. Thirteen additional exact Physical-L3 label leaks are handled identically. Six of the excluded rows previously carried HOLD, reducing the displayed count from 82 to 76. The protected Physical set remains exactly 182 cards.

A.6 Definition audit and Agentic expansion (v2.7–v2.8)

The v2.7 audit treats L3 load as a triage signal rather than evidence of semantic fit. It removes 1,129 imported hierarchy-name scaffolds from L4 definitions, applies nine evidence-constrained remaps, and marks 692 cards for decision review. Release v2.8.0 then adds four Agentic-specific L3 families: *Goal & Planning*, *Tool Calling*, *Memory*, and *Oversight & Control*. Each node has an English-Korean definition and explicit references. A non-Physical card moves only when its mechanism is intrinsically Agentic, the new family is more specific than every existing General or Agentic alternative, and the necessity gate rejects adequate coverage by the prior taxonomy. Formally, for card i and new family k,

$$k_i^* = \arg\max_{k \in K_{new}} s(i,k), \qquad \Delta_i = s(i,k_i^*) - \max_{j \in K_{old}} s(i,j), \tag{10}$$

and movement additionally requires mechanism-specific rules, sufficient absolute fit, positive separation, and no Physical lock. The deterministic iteration changes 31 paths on pass 1 and zero on pass 2. Counts in the new families are 11, 11, 3, and 6, respectively. All other placements are preserved, and the v2.8.0 review queue contains 685 cards.

A.7 Memory-card scope differentiation (v2.13.0)

A near-duplicate audit found that the three cards assigned to Agentic Memory had pairwise BGE-M3 cosines of 0.822 to 0.874, placing every pair above the 99.98th percentile of all released card pairs. In particular, RAI4-0011 and RAI4-1670 expressed a broad-narrow subsumption relationship rather than clearly distinct atomic risks. Release v2.13.0 therefore preserves all three immutable IDs, references, impact metrics, and L3 paths, while revising their intensions as follows.

Table 11: Source-scoped differentiation of the three Agentic Memory cards.

L4 ID	Revised label	Distinguishing mechanism
RAI4-0011	Agent context poisoning	Adversarial corruption of retained or retrievable summaries, embeddings, RAG entries, or shared contextual state, without requiring alteration of a dedicated long-term store
RAI4-0484	Unsafe memory accumulation	False, private, stale, or malicious records accumulate because validation, provenance, retention, or deletion is inadequate; a deliberate poisoning attack is not required
RAI4-1670	Persistent agent-memory poisoning	Adversarial insertion or modification of records in dedicated long-term, cross-session memory, with persistent effects on later retrieval, planning, and action

The revisions are editorial hypotheses subject to source-entailment review, not a declaration that the original source rows were independent risks. All three cards are therefore marked HOLD. If reviewers cannot support the stated boundaries directly from the evidence, the conservative alternative is canonical consolidation with aliases and preserved source provenance rather than further lexical renaming.

A.8 Sparse Agentic-family retirement and forced migration (v2.14.0)

The four experimental Agentic families introduced in v2.8.0 contained only 39 cards in v2.13.0: Goal & Planning (16), Tool Calling (11), Memory (3), and Oversight & Control (9). Release v2.14.0 removes these nodes from the active hierarchy without erasing their history. Their identifiers, bilingual labels, definitions, references, source counts, and introduction metadata are archived in a machine-readable retirement record; the identifiers are explicitly prohibited from reuse. All L4 identifiers, definitions, references, risk metrics, and provenance are preserved.

Every affected card is force-migrated to the most defensible active non-Physical family under a mechanism-first review. Candidate destinations are restricted to the 26 active non-Physical L3 families. BGE-M3 seed similarity is used as a diagnostic ranking, while explicit mechanism rules distinguish goal failure, unsafe execution authority, network propagation, privacy, excessive capability use, misinformation, overconfidence, and contestability. No Physical path is eligible. Because this operation is a forced operational consolidation rather than human adjudication, all 39 migrations carry `HOLD`; 28 were newly marked and 11 retained an existing marker.

Table 12: Destinations of the 39 v2.14.0 forced migrations.

Active destination L3	Cards
Goal Misalignment	15
Excessive Authority	9
Misinformation/Disinformation	4
Non-Contestability	4
Overconfidence	3
Network Effects	2
Privacy	1
Over-Extension	1

The retired families are not declared conceptually invalid. Their sparse granularity is retained as an auditable design experiment that can be restored only through a later, human-approved taxonomy decision. Release v2.14.0 has 50 active L3 nodes, while the four retired nodes remain available solely as archived lineage metadata.

A.9 Domain-specific HOLD review paths (v2.15.0)

Release v2.15.0 makes review status navigable in the hierarchy by introducing one HOLD L2 path under General and one under Agentic AI. Because L4 cards must terminate at L3, each path contains one child node, *Taxonomy Decision Hold*. The operation moves 626 General and 94 Agentic cards, for 720 cards in total. Physical AI contains no HOLD cards and its 182 locked mappings remain byte-equivalent apart from the release identifier.

This review path is orthogonal to semantic classification. For every moved card, the former L2 and L3 identifiers, bilingual labels, definitions, references, metrics, and migration metadata are preserved in `hold_semantic_path`. Accordingly, the public hierarchy has four canonical L2 display categories and 52 navigable L3 nodes: 50 semantic risk families plus two review-path nodes. The HOLD nodes must not be counted as new risk concepts or used as semantic prototypes in reliability analysis.

A.10 Evidence-enrichment remediation (v2.16.0)

Some HOLD reasons are remediable without changing the taxonomy. Label-definition mismatch, inherited taxonomy-family prose, and weak or overly broad source support are treated as data-quality defects before they are treated as final taxonomy decisions. Release v2.16.0 therefore introduces a remediation audit over all 720 HOLD cards. The audit separates cards into six operational classes: taxonomy-structure review, mapping revalidation, title-definition rewrite candidate, human taxonomy decision, evidence-enrichment candidate, and evidence plus taxonomy review.

The first-pass audit finds 268 cards that mainly expose missing or insufficient L3 structure, 200 cards requiring mapping revalidation, 147 cards suitable for title-definition rewriting, 72 cards requiring human boundary decisions, 21 cards suitable for evidence enrichment, and 12 cards requiring both evidence enrichment and taxonomy review. This confirms that weak evidence is repairable for a subset of HOLD, but it is not the dominant HOLD reason.

The first v2.16.0 enrichment pass updates 10 governance and interaction-risk cards with 11 directly relevant, online-verified references and rewrites their definitions as specific L4 mechanisms. The affected cards include risk classification error, capability evaluation non-disclosure, benchmark governance failure, red-team governance failure, regulatory arbitrage, regulatory capture, international coordination failure, cross-border enforcement gap, AI registry incompleteness, and AI ethics washing. No card is automatically removed from HOLD in this pass. Evidence enrichment is a necessary repair step; HOLD clearance still requires a follow-up mapping reliability rerun or explicit human approval.

A.11 Minimal Physical AI transfer (v2.17.1)

Release v2.17.1 applies a deliberately conservative transfer rule to the v2.17.0 HOLD queue. A card is moved under Physical AI only when its own label and definition directly describe an embodied, robotic, autonomous-vehicle, drone, cyber-physical, physical-labor, or physical-infrastructure mechanism. Cards that merely use a Physical AI example inside a broader legal, governance, warfare, or general-system discussion remain in HOLD. The protected 182-card Physical AI source set is not edited; the release only adds seven newly routed global cards to Physical paths for public review.

Table 13: Minimal HOLD-to-Physical transfers in v2.17.1.

L4 ID	Risk label	Physical AI target L3
RAI4-0520	Labour displacement	Labor Displacement
RAI4-0640	Purposeful or malicious harm	Purposeful / Malicious Harm
RAI4-0797	Misinformation generation	Misinformation
RAI4-1294	Catastrophic autonomous-weapon target risk	Purposeful / Malicious Harm
RAI4-1410	Specification gaming	Robot Control
RAI4-1549	Damage to critical infrastructure	Software Vulnerabilities and Design Flaws
RAI4-1678	Physical-world backdoor in robotic manipulation	Cybersecurity Threats

Nine physically adjacent cards are intentionally retained in HOLD because the card mechanism is broader than Physical AI or primarily legal, governance, or general-purpose in scope. This keeps the transfer minimal: HOLD decreases from 734 to 727, and Physical-path cards increase from 182 to 189 without changing the locked Physical source-card records.

Appendix B HOLD policy details

B.1 Qualitative retention criteria

A card remains or becomes `HOLD` when any of the following substantive conditions applies:

- the label and mechanism definition describe different risks, or the cited source does not directly support the stated failure mode;
- the card is multi-mechanism, the destination L3 is overloaded, or two plausible L3 paths cannot be separated without a policy judgment;
- the risk exposes a taxonomy gap, so the present L3 is merely the closest operational path rather than a definitionally adequate family;
- an Agentic destination lacks an intrinsically Agentic execution loop, tool action, persistent memory, autonomous planning, or oversight mechanism;

- a non-authoritative card appears Physical but is outside the locked 182-card Physical set, or a proposed change would disturb a locked Physical path;
- independent reviewers reject the proposal, disagree on the destination, or do not reach exact high-confidence consensus; or
- the card was moved by constrained EM and has not yet received explicit human approval.

B.2 Historical HOLD baseline (v2.12.0–v2.13.0)

The 689 v2.12.0 held cards are grouped by the stored primary reason in Table 14. Composite taxonomy-gap labels are counted once in the taxonomy-gap row.

Table 14: Primary reasons for the v2.12.0 HOLD queue.

Primary reason group	Cards
Taxonomy gap	280
Anthropomorphism mechanism not established	173
Overloaded L3 or low evidence fit	138
Expert review without consensus	28
Constrained-EM remap pending review	20
Physical-boundary conflict	16
Expert rejection	16
Expert disagreement	7
Multiple mechanisms	5
Other named exceptions	4
Low absolute fit	2
Total	689

Release v2.13.0 adds three review markers without changing any L3 path. Two cards require adjudication of a semantic near-duplicate scope boundary, and one additionally requires direct source-entailment review. The current queue therefore contains 692 cards (40.4%).

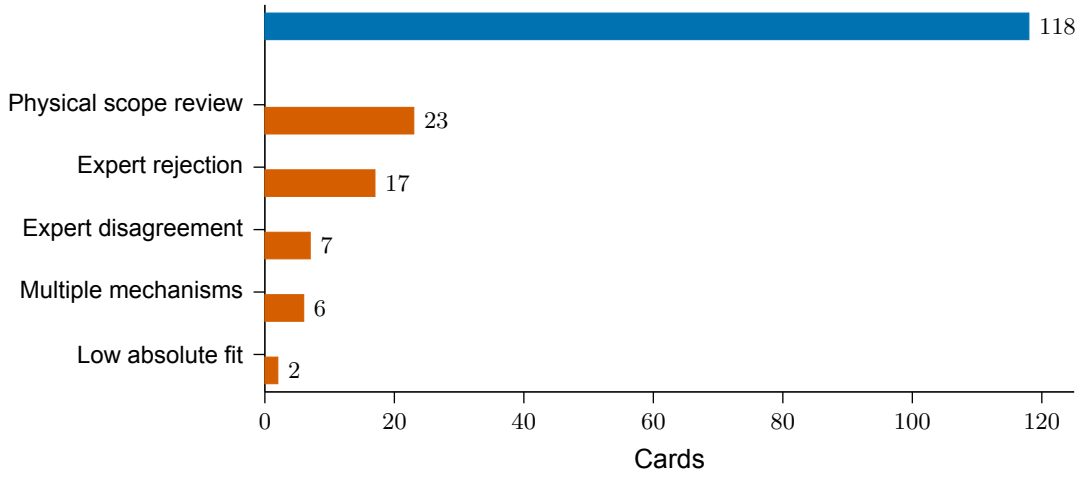

Figure 8: **Disposition of the 173 Stage 2 holds.** The 55 difficult cases (vermillion) are force-assigned and marked HOLD; 118 suitability-reserve cards (blue) are also force-assigned.

Appendix C Exploratory analyses

C.1 Candidate L3 families for General and Agentic AI

The v2.17.1 taxonomy does not add new General or Agentic L3 families. However, the HOLD reason analysis indicates where future L3 additions are most likely to improve semantic reliability. The diagnostic rule is conservative: a new L3 is considered only when it explains a recurring HOLD cluster, has a mechanism that is not already captured by an existing L3 definition, and is likely to reduce overloaded or ambiguous assignments without fragmenting the taxonomy. The analysis is therefore a design hypothesis for the next review round, not a published mapping change.

General AI has the larger improvement opportunity. Of the v2.17.0 HOLD queue, 280 cards are tagged as taxonomy-gap or missing-scope cases, 173 as Anthropomorphism mechanism ambiguity, and 138 as overloaded-L3 or weak evidence-fit cases. Keyword and definition review further shows recurring General-side clusters around governance, transparency, security, dependency, resources, labor, and power. These clusters are semantically broader than single L4 labels but narrower than the current L1 or L2 structure, making them plausible L3 candidates.

Table 15: Conservative candidate L3 families for future General AI review.

Priority	Candidate L3	Korean label	Expected reliability effect
1	Governance and Accountability	거버넌스·책임성	Reduces governance and audit taxonomy gaps
2	Transparency and Explainability	투명성·설명가능성	Separates documentation, opacity, and explanation risks
3	AI Security	AI 보안	Separates non-Physical security, attack, and vulnerability risks
4	Dependency and Manipulation	의존·조작	Relieves Anthropomorphism over-assignment
5	Resource and Environmental Harm	자원·환경 영향	Captures compute, energy, water, and carbon harms
6	Labor and Power	노동·권력 불균형	Captures labor, concentration, and inequality mechanisms

The first four General candidates are the strongest. *Governance and Accountability* directly addresses the largest taxonomy-gap theme, including audit, regulation, certification, reporting, accountability, and liability cards that are currently forced into less specific safety families. *Transparency and Explainability* separates model-card, system-card, documentation, opacity, and contestability risks from general interaction or system-safety families. *AI Security* provides a non-Physical home for prompt injection, poisoning, jailbreak, exfiltration, backdoor, and software vulnerability risks when the risk mechanism is not specifically embodied. *Dependency and Manipulation* is expected to improve the overloaded Anthropomorphism region by moving trust, persuasion, overreliance, parasocial-bonding, and behavioral-manipulation cards whose operative mechanism is dependence or influence rather than human-like presentation.

Agentic AI should be treated more conservatively. The prior v2.8.0 expansion tested four intuitive Agentic families: Goal and Planning, Tool Calling, Memory, and Oversight and Control. Those families were retired in v2.14.0 because the released L4 counts were sparse and the resulting separation was not stable enough to justify permanent L3 expansion. The current evidence therefore does not support recreating those four categories as published L3 families. Two narrower Agentic candidates may still be worth an experimental ablation because they reflect agent-specific failure modes rather than generic General AI risks.

The recommended next experiment is therefore minimal: test the four strongest General candidates first, then run an ablation with zero, one, or two Agentic candidates. A candidate should be accepted only if it improves non-HOLD top-1 containment, top-5 containment, median assignment margin, negative-margin share, and perturbation stability without creating a new sparse or unstable family. Until that criterion is met, the current v2.17.1 release should remain unchanged.

Table 16: Candidate Agentic AI L3 families for experimental ablation only.

Priority	Candidate L3	Korean label	Reason for caution
1	Tool and Environment Security	도구·환경 보안	Agent-specific, but overlaps with General AI Security
2	Traceability and Audit	추적성·감사	Agent-specific action provenance, but overlaps with governance

C.2 Counterfactual reliability simulation

The candidate L3 proposal was tested as a counterfactual simulation rather than as a release change. The simulation uses the same pinned BGE-M3 card embeddings as the v2.17.0 audit and evaluates v2.17.1 under three conditions: the current 50 semantic L3 families, the current taxonomy plus four General AI candidates, and the current taxonomy plus four General AI candidates and two Agentic AI candidates. Physical source cards are locked. A card can move to a candidate family only when its English title or definition matches the candidate keyword gate and its BGE-M3 similarity to the candidate definition exceeds its similarity to the current L3 seed by at least 0.012. This is a conservative automatic gate; it is not a human gold-standard adjudication.

Table 17: Counterfactual all-card reliability under candidate L3 expansion.

Condition	Families	Moved	Top-1	Top-5	Neg. margin	Noise 0.05
Current v2.17.1	50	0	64.6%	92.9%	35.4%	72.3%
General +4	54	203	63.1%	91.9%	36.9%	71.9%
General +4, Agentic +2	56	243	63.2%	91.6%	36.8%	72.1%

Table 18: Counterfactual non-HOLD reliability under candidate L3 expansion.

Condition	Families	Moved	Top-1	Top-5	Neg. margin	Noise 0.05
Current v2.17.1	50	0	76.4%	95.5%	23.6%	78.2%
General +4	54	42	75.9%	95.9%	24.1%	77.6%
General +4, Agentic +2	56	68	75.8%	95.7%	24.2%	77.7%

The simulation does not support immediate L3 expansion. The four General candidates move 203 cards, including 161 held cards, but all-card top-1 containment falls from 64.6% to 63.1%, top-5 containment falls from 92.9% to 91.9%, negative-margin share rises from 35.4% to 36.9%, and $\sigma = 0.05$ perturbation agreement falls from 72.3% to 71.9%. Adding the two Agentic candidates moves 243 cards in total but still fails to improve the main reliability criteria. Non-HOLD results show the same pattern: the candidate expansions create only marginal top-5 changes while reducing top-1 containment, margin quality, and perturbation stability.

The only metric that improves under the all-card counterfactual is EM-to-release agreement, rising from 20.6% to 23.4% for General +4 and 23.3% for General +4 plus Agentic +2. This improvement is not sufficient for release adoption because it is accompanied by weaker top-k containment and lower stability. The practical conclusion is that the proposed General and Agentic L3 candidates remain useful review hypotheses, but they should not be added to the public taxonomy until a human-adjudicated gold set and a stricter ablation show simultaneous improvement in top-1 containment, top-5 containment, median margin, negative-margin share, and perturbation stability.

C.3 Two-dimensional release-state vector space

Figure 10 projects all 1,711 v2.17.1 L4 cards into a two-dimensional vector space. The input vectors are the pinned BGE-M3 card embeddings used in the v2.17.0 audit; v2.17.1 changes only seven routing decisions and does not rewrite card text, so the same embedding matrix is valid for this diagnostic. The projection

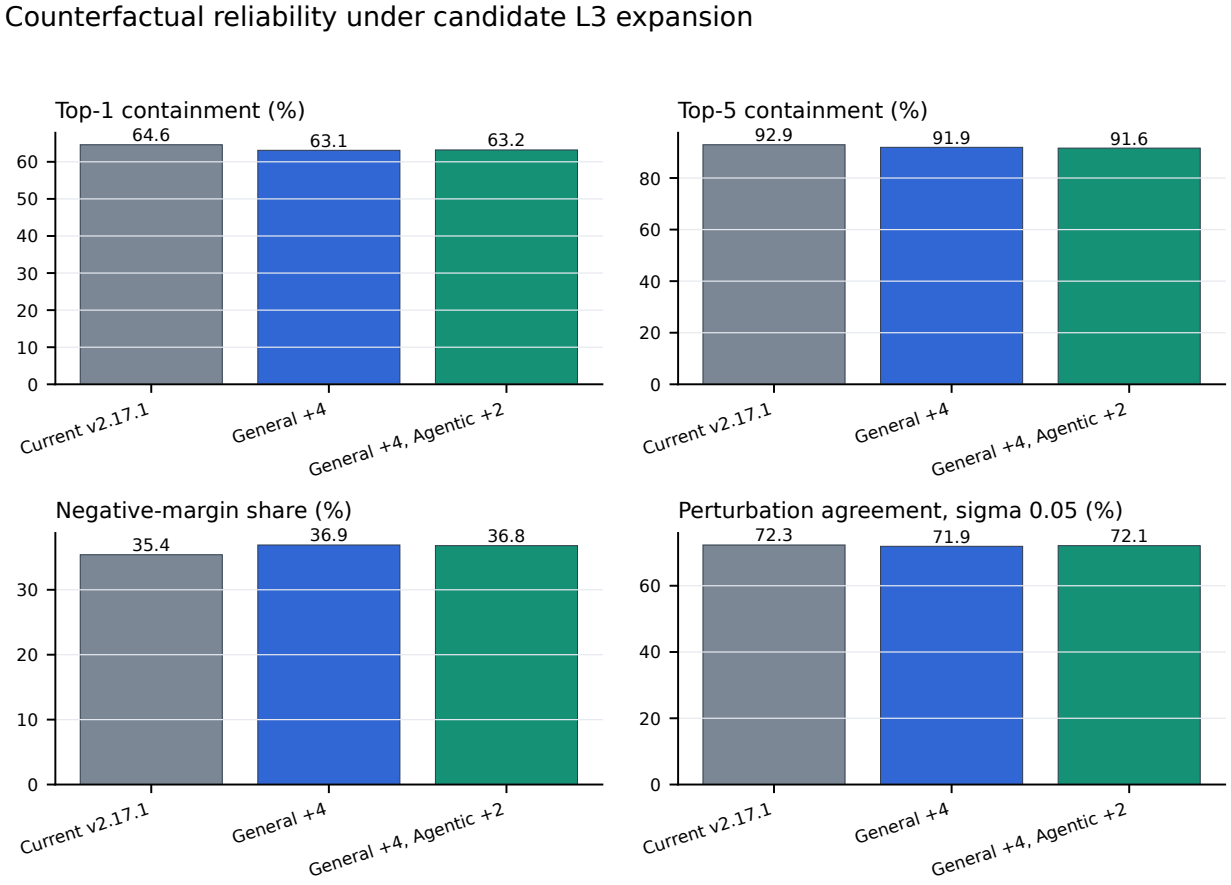

Figure 9: **Counterfactual reliability under candidate L3 expansion.** Bars compare the current v2.17.1 release with two candidate-expansion conditions. Higher is better for top-1, top-5, and perturbation agreement; lower is better for negative-margin share. Candidate expansion creates plausible movement candidates but does not improve the main reliability criteria.

uses PCA on L2-normalized embeddings. The first two components explain 5.0% and 3.8% of variance, respectively. The figure should therefore be read as a low-dimensional diagnostic view of release state, not as proof of discrete taxonomy clusters.

Appendix D Historical reliability results

D.1 Full-release results before HOLD separation (v2.12.0–v2.14.0)

Table 19: Released-centroid support for the v2.14.0 forced migrations.

Subset	Cards	Mean cosine	Median margin	Top-1	Top-3
Migrated v2.14.0	39	0.746	-0.005	46.2%	71.8%
All other cards	1,672	0.769	0.013	65.1%	86.4%

The migrated subset is weaker than the unchanged population: 53.8% of its cards have negative assignment margins, and its top-1 containment is only 46.2%. This result supports the conservative policy decision to retain HOLD on every forced migration. It does not invalidate the operational destinations; it identifies them as priority items for human adjudication.

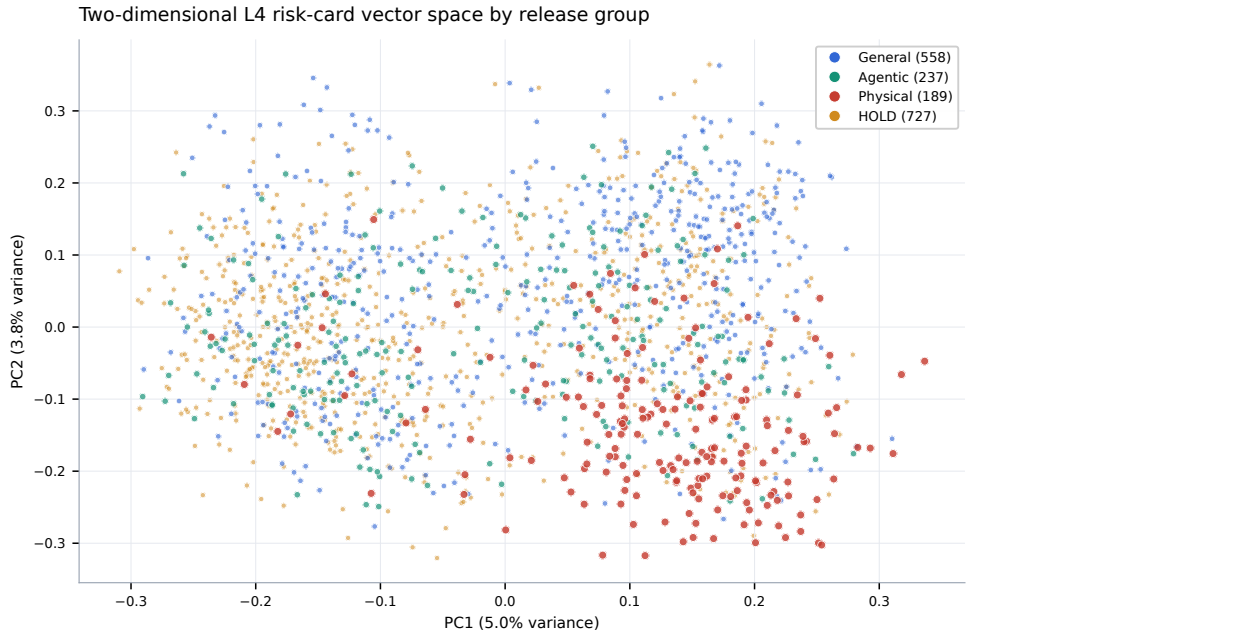

Figure 10: **Two-dimensional L4 risk-card vector space by release group.** Each point is one L4 card. Colors show the public release group after v2.17.1: General AI, Agentic AI, Physical AI, or HOLD. HOLD is a review state rather than a semantic family. Physical includes the 182 locked Physical source cards plus seven conservative v2.17.1 transfers. The overlap between HOLD and the three semantic domains supports the current interpretation of HOLD as a cross-cutting adjudication queue rather than a separate risk domain.

D.2 Reliability after excluding HOLD review paths (v2.15.0)

The second analysis uses the identical encoder, seed, centroid procedure, 5,000 matched-size permutations, and 200 perturbation repeats, but excludes all 720 HOLD cards before constructing family centroids. It assesses 991 non-HOLD cards across the same 50 semantic L3 families; none of those families is empty. The two HOLD review nodes are navigation constructs and are not embedded or treated as risk-family prototypes.

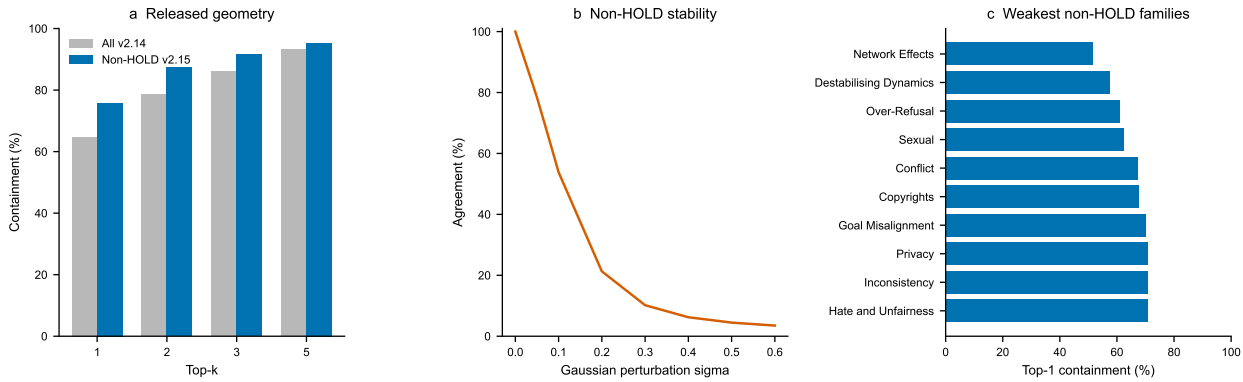

Figure 11: **Reliability after HOLD separation (v2.15.0).** (a) Top-k containment for the 991 non-HOLD cards is shown against the full v2.14.0 population. (b) Perturbation stability is calculated only on non-HOLD centroids and cards. (c) The ten weakest non-HOLD families identify the next review priorities.

Table 20: Reliability with HOLD embedded and after HOLD exclusion.

Metric	All v2.14.0	Non-HOLD v2.15.0
Cards assessed	1,711	991
Semantic L3 families	50	50
Mean within-family cosine	0.768	0.779
Median assignment margin	0.012	0.024
Negative-margin share	35.3%	24.2%
Top-1 containment	64.7%	75.8%
Top-2 containment	78.6%	87.3%
Top-3 containment	86.1%	91.5%
Top-5 containment	93.1%	95.4%
EM agreement with released paths	21.8%	31.7%
Agreement at $\sigma = 0.01$	94.5%	95.6%
Agreement at $\sigma = 0.05$	72.8%	78.4%
Permutation p-value	0.0002	0.0002

The conditional non-HOLD subset is more coherent: top-1 containment rises by 11.1 percentage points, the median margin approximately doubles, and the negative-margin share falls by 11.1 points. EM reaches a fixed point after 13 iterations and agrees with 31.7% of released paths (adjusted Rand index 0.187; normalized mutual information 0.501). Observed cohesion is 0.779 versus a matched-size null mean of 0.742, with zero exceedances and plus-one $p = 0.0002$.

These are not before and after accuracy estimates on the same population. The later analysis deliberately removes the review-priority cases that were known to have weaker fit, changes every centroid, and reduces the sample from 1,711 to 991. The improvement therefore validates HOLD as an effective triage separation signal; it does not prove that every non-HOLD assignment is correct or that moving a card into HOLD improves its semantic classification.

D.3 HOLD sensitivity analysis (v2.15.0, historical)

We next compare the same three diagnostics under two explicit conditions: *HOLD included*, where all 1,711 v2.14.0 cards remain embedded in their semantic L3 assignments, and *HOLD excluded*, where the 720 review-path cards are removed and only the 991 non-HOLD cards are assessed. Both conditions use BGE-M3 embeddings, seed 20260721, the same 50 semantic L3 families, 5,000 matched-size label permutations, and 200 perturbation repeats. The two HOLD navigation nodes introduced in v2.15.0 are excluded from centroid construction.

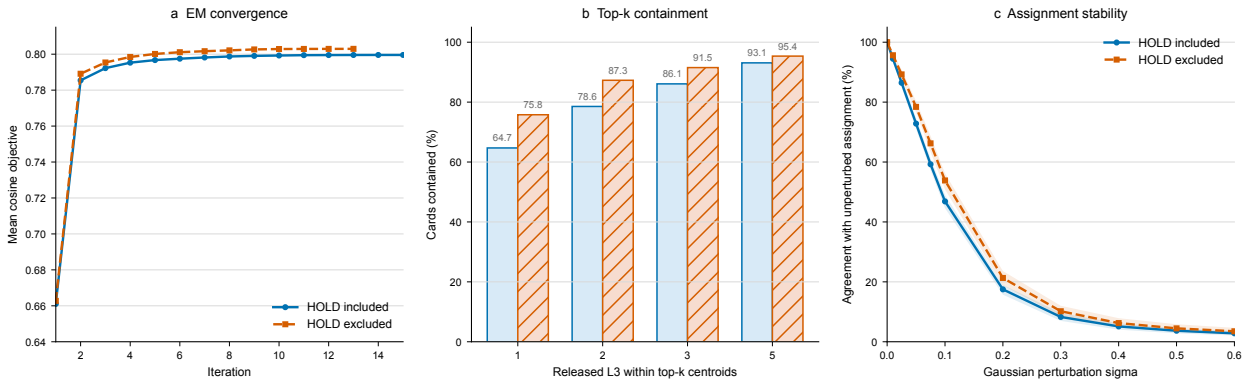

Figure 12: **HOLD sensitivity analysis.** (a) Seed-initialized spherical EM reaches a fixed point after 15 iterations with HOLD included and after 13 iterations with HOLD excluded. (b) Released-L3 top-k containment improves across $k = 1, 2, 3, 5$ after removing HOLD cards. (c) Assignment stability under Gaussian embedding perturbation is higher on the non-HOLD subset across the tested noise schedule. Shaded bands indicate 95% intervals across 200 perturbation repeats.

Table 21: Sensitivity results with HOLD included and excluded.

Diagnostic	HOLD included	HOLD excluded
Cards assessed	1,711	991
HOLD cards excluded	0	720
EM iterations to fixed point	15	13
Final EM objective	0.800	0.803
EM agreement with released paths	21.8%	31.7%
Top-1 containment of released L3	64.7%	75.8%
Top-2 containment of released L3	78.6%	87.3%
Top-3 containment of released L3	86.1%	91.5%
Top-5 containment of released L3	93.1%	95.4%
Perturbation agreement, $\sigma = 0.01$	94.5%	95.6%
Perturbation agreement, $\sigma = 0.05$	72.8%	78.4%

The sensitivity result is directionally consistent across the three requested tests. HOLD exclusion shortens EM convergence by two iterations, raises final objective from 0.800 to 0.803, increases released-path agreement by 9.9 percentage points, raises top-1 containment by 11.1 points, and raises $\sigma = 0.05$ perturbation agreement by 5.6 points. The correct interpretation is not that the same cards became more reliable. Rather, the HOLD policy successfully concentrates low-fit and taxonomically ambiguous cases, leaving a cleaner subset whose released L3 assignments are more stable under the tested embedding geometry.

Appendix E Rounding and reporting policy

All counts are reported as integers. Percentages below 1 are rounded at the third decimal place and displayed to two decimal places; percentages with an integer component of at least 1 are rounded and displayed to one decimal place. For example, the L4 registry represents 0.04% of the integrated AI evidence count, the AI-risk view represents 2.1%, final L3 coverage is 100.0%, and the decision-required share is 42.1%. Exact floating-point ratios remain available in the machine-readable statistics artifact.

This document is the journal edition of RAI Risk Taxonomy Technical Report 2.0 at release state v2.17.2.