Technical Report

DRAFT — for review, not for citation

A Hybrid Corpus-to-Taxonomy Pipeline for Physical AI Risk Cards

Data Collection, Processing, Deduplication, and Hierarchical Assignment

Report type	Technical Report
Report title	A Hybrid Corpus-to-Taxonomy Pipeline for Physical AI Risk Cards
Report version	v1.0 — **Draft (for review)**
Taxonomy release	Responsible AI Risk Taxonomy v2.0, Physical AI Risk Taxonomy slice
Author	Youngsam Chun
Affiliation	Responsible AI Dept., Frontier AI Lab
Contact	ys.chun@kt.com
Local data snapshot	Verified on 2026-07-04
Primary artifacts	HTML taxonomy, CSV/JSON card exports, reference map, methodology PDF
Reproducibility status	Local source queries, file-level denominators, algorithms, and release checks documented

Responsible AI Dept., Frontier AI Lab

July 2026

Executive Summary

This manuscript documents the data collection, processing, analysis, and risk-card generation procedure used to construct the Physical AI slice of the Responsible AI Risk Taxonomy v2.0. The workflow follows a hybrid taxonomy design: high-level structure is anchored in authoritative safety, risk-management, and embodied-AI references, while L4 risk cards are generated bottom-up from a multi-source corpus of scientific publications, policy reports, and patents. The updated local integrated document space contains 3,906,767 AI records, 741,428 AI infrastructure records, 383,149 Physical AI records, and 82,971 Physical AI risk records after within-group deduplication across scientific publications, policy reports, and patents. The released local taxonomy snapshot contains 182 L4 risk cards, 24 L3 sub-categories, 3 L2 categories, and 360 card-level evidence/reference rows. We formalize two procedures: (i) risk-candidate generation with semantic deduplication and (ii) EM-based L3 family assignment.

Keywords: Physical AI; AI risk taxonomy; robotics safety; embodied AI; cyber-physical systems; semantic deduplication; BGE-M3; HDBSCAN; reproducible technical report.

Reproducibility statement. This report is written as a technical reconstruction protocol. It specifies the source families, retrieval artifacts, local file denominators, filtering thresholds, embedding inputs, clustering and deduplication procedures, hierarchy assignment procedure, validation checks, and release artifacts required to reproduce the Physical AI Risk Taxonomy construction workflow.

Contents

1 Scope and Design Rationale

Physical AI is defined here as AI deployed in systems that sense, reason, and act in the physical world, including robots, humanoids, drones, autonomous vehicles, cyber-physical systems, and embodied agents. This scope follows recent Physical AI and embodied-safety work that treats world models, robot constitutions, and embodied benchmarks as distinct from text-only AI risk settings [1, 2, 3]. The taxonomy is organized as:

$$\begin{aligned} L0 \text{ Ownership Levels} \rightarrow L1 \text{ Physical AI Risks} \\ \rightarrow L2 \text{ Categories} \rightarrow L3 \text{ Sub-categories} \rightarrow L4 \text{ Risk Cards.} \end{aligned} \tag{1}$$

The L2 layer distinguishes the locus of risk responsibility: internal system safety, interaction safety at the human/environment interface, and broader societal safety. The L4 layer is the operational unit: each card has a stable identifier, bilingual label and definition, severity and probability proxies, short evidence justifications, references, and 3H1R alignment tags.

The methodological problem is that Physical AI risk taxonomies cannot be built purely top-down or purely bottom-up. A top-down taxonomy is stable and institutionally legible, but may miss emerging embodied, agentic, and cyber-physical failure modes [4, 5]. A bottom-up corpus pipeline captures emerging risks, but can produce noisy, overlapping, or non-actionable categories. The adopted design therefore combines: (1) theory-anchored upper levels, (2) corpus-driven L4 candidate discovery, (3) semantic deduplication, and (4) expert review for boundary cases. This hybrid stance is consistent with prior language-model risk taxonomies and AI risk repository work, which combine expert structure with systematic evidence aggregation [6, 5].

2 Background: From Asimov's Laws to Alignment-Grounded Risk Cards

The problem of constraining a physical machine's behavior did not begin with reinforcement learning; its first serious formulation is literary. Asimov's Three Laws of Robotics, together with the later Zeroth Law protecting humanity as a whole, embedded a few ranked, general-purpose priorities in the machine itself. They ask the machine to not injure a human or through inaction allow harm, to obey human orders except where they conflict with the first law, and to protect its own existence except where that conflicts with the first two [7]. The template endures, and so do its limitations. Stated in natural language, the laws are ambiguous exactly where it matters. The terms "harm," "human," and "inaction" have no operational definition a controller could evaluate, their ranking resolves conflicts only when the terms are already crisp, and they carry no representation of the machine's capability limits or of the chance that its perception is wrong. Murphy and Woods put the critique on a formal footing, arguing that the Laws are an unworkable basis for real robotics and relocating responsibility onto the humans who design and deploy the system rather than onto the robot [8]. The lesson we follow through this section is that naming priorities does not tell you how to apply them to a specific act under uncertainty.

One influential response to the brittleness of hand-specified behavior was to *learn* alignment from human feedback rather than legislate it in rules. Deep reinforcement learning from human preferences established the preference-comparison mechanism that modern reinforcement learning from human feedback (RLHF) builds on, communicating goals through pairwise human judgments without a hand-coded reward [9]; instruction tuning with human feedback then scaled this to large language models (LLMs), producing assistants that follow user intent and, on the authors' evaluations, reduce toxic output [10]. This is the practical machinery behind today's helpful–harmless–honest assistants.

Learning the objective, however, does not escape the specification problem; it inherits it. Russell frames the fixed-objective "standard model" of AI as the root difficulty, since a machine that optimizes a given objective perfectly will pursue it past the point its designers intended [11]. Agents duly satisfy the letter

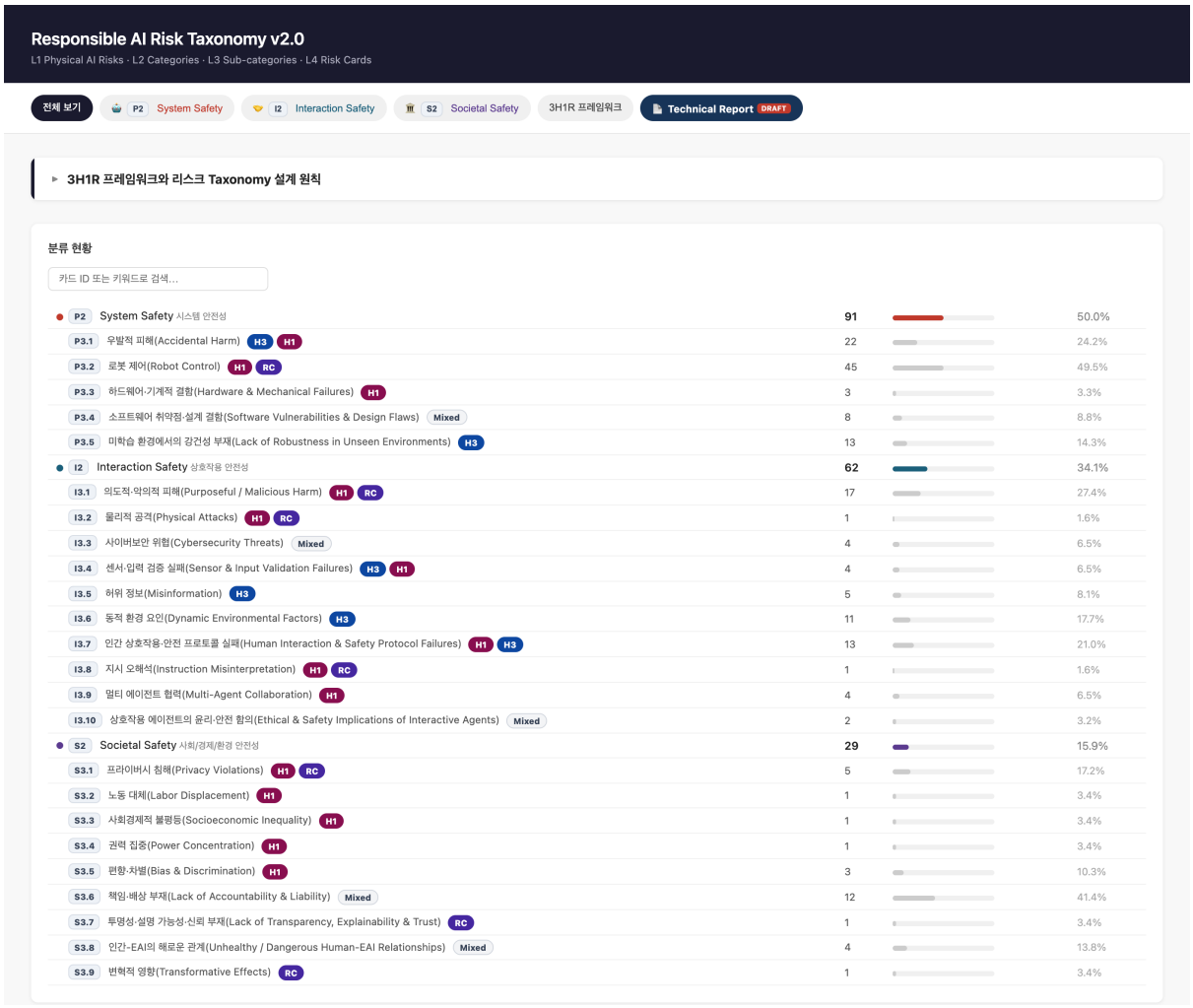

Figure 1: Released taxonomy interface

Note: The interface is used as the visual inspection layer for the Physical AI Risk Taxonomy. It presents the L0–L4 hierarchy, L2 source categories, L3 sub-category distribution, L4 risk-card counts, and representative 3H1R badges. The screenshot is included as a visual artifact of the release state; the report does not rely on the web page as a bibliographic source.

of a *learned* reward while violating its intent [12], and a systematic survey documents RLHF's open problems, which range from misspecified and inconsistent human feedback to reward-model misgeneralization and the limited scalability of human oversight [13]. A second limit is structural. RLHF as used to align language models operates over *text*, so it does not, by itself, represent embodied action, irreversible physical consequence, or the authority a machine holds. A model can produce a perfectly polite, "harmless"-sounding plan and still execute a motion that exceeds a joint or payload limit, grasp a blade by its edge, or compose two individually safe motions into a collision, all failures a transcript-level reward cannot see [2, 3].

Principle-based alignment sits *on top of* feedback rather than replacing it, since a few written principles supply the targets that optimization aims at. The helpful–harmless–honest desiderata [14] and constitutional training, which generates feedback from a written constitution and so reduces reliance on human preference labels [15], operationalize this for language models. Physical AI needs one axis more. We extend the three desiderata with an explicit *Role Consistency* (RC) dimension, meaning behavior that stays within the authority and control boundary granted to the machine, and call the combined set **3H1R**. Role consistency is added rather than folded into Harmlessness because authority violations occur even when no physical harm results; it is the operational descendant of Asimov's obedience law.

A parallel governance line moved control from the action to the model. Capability-threshold frameworks gate deployment on whether a system crosses a dangerous-capability tier [16, 17, 18]. This is the right instrument for frontier risk but coarse for Physical AI, where the failures that injure people are per-action and per-context rather than properties of a tier, since a well-below-threshold policy can still crush a hand through a mundane control error. Functional-safety engineering supplies per-failure rigor. The safety of the intended functionality (SOTIF) addresses hazards from performance limitations and foreseeable misuse absent any fault, and goal-based safety cases structure the argument for autonomous products [19, 20]. Yet both were shaped around engineered, specifiable systems and map awkwardly onto learned policies whose failures are emergent and semantic rather than enumerable at design time [21]. The taxonomy literature is itself split. Top-down expert taxonomies give stable structure but lag the field and reflect their authors' priors [6, 22], while bottom-up corpus mining captures breadth and currency but yields flat, overlapping inventories without a defensible hierarchy [5]. Neither suffices alone.

Across all of these paradigms the same seam reopens. Rules name priorities but cannot apply them under uncertainty; learned rewards inherit specification gaming and, as deployed, see only text; written principles, 3H1R included, say *which* values should hold but not *how* to hold a specific physical action to them, that is, how to evaluate the behavior against a value, trace it to evidence, estimate its severity, and record it as a reusable unit. Emerging embodied-safety benchmarks supply valuable test items but not that unit [2, 3]. We supply it directly, through two commitments that organize the rest of this report. The atomic unit is the **risk card**, a single identifiable failure mode with an operational definition, supporting evidence, severity and probability estimates, and 3H1R tags, auditable in a way a natural-language law or a scalar reward is not. The construction is **hybrid**, pairing an expert-defined upper structure with corpus-derived candidates, because an expert-only taxonomy reproduces the a-priori blind spots that undid the Three Laws, while a corpus-only taxonomy inherits the "whatever the data contains" misgeneralization of learned objectives. Reconciling the two is the method-level form of the values-versus-evidence tension this section has tracked throughout.

Contributions

Theoretical. We recast the helpful–harmless–honest desiderata as an operational lens for embodied action and add an explicit Role Consistency axis (3H1R) that captures authority- and control-boundary failures invisible to a harm-only reading and absent in text-only settings. We give a constrained-optimization account of the Harmless–Helpful trade-off (Section 16): modeling an embodied policy as a constrained Markov decision process [23, 21], efficiency pressure predictably drives behavior toward the high-capability,

high-harm quadrant, making the taxonomy's most dangerous cell a consequence of how these systems are optimized rather than an anomaly.

Methodological. A hybrid corpus-to-taxonomy pipeline reconciles a human-defined upper structure (L0–L2 and 24 human-defined L3 families) with bottom-up candidates discovered by BGE-M3 embedding and HDBSCAN clustering over papers, policy reports, and patents; each of the 182 L4 cards is reference-grounded and auditable at the level of the individual failure mode; and an EM-style hierarchy assignment places cards under L3 with reported reproducibility and robustness checks, so the hierarchy is measured rather than asserted.

Practical. We release an auditable 182-card catalogue with bilingual (Korean/English) coverage and 3H1R tags, usable as a governance checklist, an evaluation and red-teaming reference, and a benchmarking scaffold for Physical AI, connected to emerging embodied-safety benchmarks [2, 3].

3 Prior Methodological Basis

The source- and institution-level measures used later in this report follow four established lines of work. Bibliometric and science-mapping methods read publication metadata to describe research fields [24, 25, 26], while responsible-metrics guidance cautions that organization counts should inform expert judgment rather than serve as quality rankings [27]. We therefore treat affiliation counts and PATSTAT[1] applicant counts [28, 29] as a provenance and credibility diagnostic of which universities, firms, agencies, and standards bodies appear in the corpus, and for policy and gray-literature sources we keep the issuing institutions and give more weight to high-authority bodies such as NIST, OECD, and ISO [30, 31], leaving quality judgments to the risk-card evidence review.

Prior work on constructing AI, Physical-AI, and agentic-AI risk taxonomies

Beyond bibliometric grounding, this report builds directly on a mature line of work that turns the scholarly and policy literature into structured risk *categories*. The closest methodological precedent is the MIT AI Risk Repository [5], which applies best-fit framework synthesis to 43 existing risk frameworks, codes roughly 1,600 risk mentions into about 700 risks, and releases a two-level domain taxonomy (seven domains and twenty-four subdomains) alongside a causal taxonomy. This literature-to-taxonomy design, in which evidence-derived candidate risks are reconciled against an expert upper structure, is the same principle this report follows for L4 discovery and L3 assignment. Comparable literature-synthesis taxonomies include the DeepMind taxonomy of language-model risks (six areas, twenty-one risks) [6], the Center for AI Safety overview of catastrophic risks (four source-based categories) [22], and the Stanford foundation-model risk survey [32].

A second group derives categories analytically or through expert policy consensus rather than corpus coding. This group includes the classic accident taxonomy of Amodei et al. [12], the pathways matrix of Yampolskiy [33], the existential-safety delegation framework of Critch and Krueger [34], the OECD framework for classifying AI systems [35] and its incident-terminology work [36], the NIST AI Risk Management Framework [37], the MITRE ATLAS adversarial-threat knowledge base [38], and the extreme-risk evaluation and advanced-assistant analyses from DeepMind [39, 40]. A third group of frontier-lab governance frameworks, such as the OpenAI Preparedness Framework [16], the Anthropic Responsible Scaling Policy [17], the DeepMind Frontier Safety Framework [18], and constitutional harm principles [15], organizes risk by capability thresholds and informs the 3H1R marking used here. For the physical and agentic setting specifically, this report draws on autonomous-system safety standards [19, 20] and on the agentic-AI risk and governance literature [4, 41, 42, 43]. Table 1 summarizes

[1] PATSTAT: the EPO Worldwide Patent Statistical Database, the primary patent-statistics source used in this report.

these works together with the actual top-level categories each released, which situates the present 3-level (L1–L3) × 182-card (L4) taxonomy within established practice.

Table 1: Prior risk-taxonomy work and its released top-level categories

Reference	Institution	Construction method	Released top-level categories (size)
A. Literature / bibliographic synthesis taxonomies			
Slattery et al. 2024 [5]	MIT	Meta-review of 43 frameworks (∼700 coded risks)	Causal taxonomy + domain taxonomy (7 domains × 24 subdomains)
Weidinger et al. 2021 [6]	DeepMind	Structured review of LM harms	Discrimination; information hazards; misinformation; malicious uses; HCI; environmental/socioeconomic (6 areas, 21 risks)
Hendrycks et al. 2023 [22]	CAIS	Narrative synthesis by source	Malicious use; AI race; organizational; rogue AIs (4)
Bommasani et al. 2021 [32]	Stanford CRFM	Foundation-model risk survey	7 thematic areas (fairness, misuse, environment, legality, economics, security)
B. Analytic / policy classification frameworks			
Amodei et al. 2016 [12]	Google Brain / OpenAI	Accident classes by failure source	Side effects; reward hacking; scalable oversight; safe exploration; distributional shift (5)
Yampolskiy 2016 [33]	U. Louisville	Two-axis analytic matrix	Timing (pre/post-deployment) × cause
Critch & Krueger 2020 [34]	UC Berkeley CHAI	Delegation-structure typology	Prepotent-AI types; single/multi principal × agent (29 directions)
OECD 2022 [35]	OECD	Policy classification of AI systems	People & planet; economy; data; model; task/output (5 dimensions)
NIST 2023 [37]	NIST	Consensus framework	7 trustworthiness traits; GOVERN/MAP/MEASURE/MANAGE
MITRE ATLAS [38]	MITRE	ATT&CK-style attack knowledge base	Adversarial tactics matrix (∼16 tactics, ∼84 techniques)
OECD 2024 [36]	OECD	Incident-terminology consensus	Incident vs. hazard; harm types (physical. . .psychological)
Shevlane et al. 2023 [39]	DeepMind	Extreme-risk evaluation	Dangerous-capability + alignment evaluations (2 axes)
Gabriel et al. 2024 [40]	DeepMind	Advanced-assistant monograph	Value alignment; safety; misuse; manipulation; well-being (thematic)
C. Frontier-lab capability-threshold governance frameworks			
OpenAI 2025 [16]	OpenAI	Capability thresholds (v2)	Biological/chemical; cybersecurity; AI self-improvement (3)
Anthropic 2023–25 [17]	Anthropic	AI Safety Levels (biosafety analogue)	ASL-1–4+ tiers (CBRN uplift, autonomy)
DeepMind 2025 [18]	DeepMind	Critical Capability Levels (v2)	CBRN; cybersecurity; ML R&D; deceptive alignment (4)
Bai et al. 2022 [15]	Anthropic	Written constitution + RLAIF	Helpful, harmless, honest via ∼16 principles
D. Physical-AI / autonomous-system risk taxonomies			
ISO 21448:2022 [19]	ISO	SOTIF hazard analysis	Functional insufficiencies; triggering conditions; foreseeable misuse (4 areas)

continued on next page

Table 1 (continued)

Reference	Institution	Construction method	Released top-level categories (size)
ANSI/UL 4600 [20]	UL	Goal-based safety case	Claim-based argument over ∼10 lifecycle topics
E. Agentic-AI risk / governance taxonomies			
Chan et al. 2023 [4]	Mila / Cambridge	Agency-feature analysis	Underspecification; directness; goal-directedness; long-term planning (4)
Shavit et al. 2023 [41]	OpenAI	Lifecycle governance practices	7 practices (task-suitability . . . interruptibility)
Chan et al. 2024 [42]	GovAI / Mila	Agent-visibility measures	Agent identifiers; real-time monitoring; activity logging (3)
Benton et al. 2024 [43]	Anthropic	Sabotage threat model + evals	Human-decision; code; sandbagging; undermining oversight (4)

Abbreviations: CBRN, chemical/biological/radiological/nuclear; ASL, AI Safety Levels; RLAIF, reinforcement learning from AI feedback; SOTIF, safety of the intended functionality; ML R&D, machine-learning research and development.

4 Design Requirements and Release Controls

This Technical Report operationalizes a hybrid Physical AI risk-taxonomy design: the upper levels (L0–L3) are stabilized through top-down governance logic, while L4 risk cards are generated from a bottom-up evidence corpus and then mapped back to the hierarchy. Table 2 records how each design requirement is implemented and where it is substantiated in this report.

Two threshold choices are treated as explicit release controls. First, the 0.52 Physical AI similarity threshold is paired with lexical overrides and audit inspection of below-threshold domains dominated by out-of-scope records such as nursing, agriculture, tourism, and general medical documents. Second, the 0.94 merge threshold is treated as an empirical separation point in the candidate-similarity distribution. It is intentionally conservative: a lower threshold would collapse distinct Physical AI failure modes, while a higher threshold would leave near-duplicate risk cards unresolved.

5 Workflow Overview

Figure 2 summarizes the end-to-end workflow from raw source collection to the released taxonomy artifacts. The pipeline is hybrid: corpus-derived evidence generates and deduplicates L4 risk candidates, while authoritative references and human review stabilize the hierarchy, definitions, and evidence justifications.

For reproducibility, each stage is documented as an auditable transformation rather than as a narrative judgment. The raw retrieval layer is preserved through the exact PATSTAT SQL and Web of Science topic query reproduced in Appendices A and B. The intermediate layer records source-family counts, local file names, row counts, column counts, and file sizes. The modeling layer specifies the embedding input text, clustering method, c-TF-IDF label extraction, reference-grounding rule, merge threshold, and L3 assignment rule. The release layer is exported from the final HTML into CSV/JSON files so that card counts, references, hierarchy assignments, and 3H1R tags can be checked independently from the rendered web page.

Table 2: Design requirements and their substantiation

Design requirement	Where it is substantiated
Taxonomy design: near-MECE, standard-aligned terms, hybrid top-down/bottom-up construction.	Scope, prior-basis, and workflow sections define the L0–L4 hierarchy, 3H1R rules, and an auditable workflow separating collection, clustering, reference grounding, review, and release.
Hierarchy: fix L0–L3 before L4 mapping (L1 Physical AI, three L2 categories, 24 L3).	Scope and released-hierarchy sections document L0–L4 structure, L2/L3 coverage, and the rule that every L3 keeps ≥ 1 L4 after review.
Source collection: build the AI corpus from Web of Science (WoS), policy reports, and PATSTAT; select Physical AI via BGE-M3 anchors plus explicit terms; keep risk records by safety/failure terms.	Data-sources section reports the integrated corpus (Table 3); Appendices A–C give the PATSTAT SQL, Physical AI/risk filters, and the Web of Science query.
L4 discovery: HDBSCAN clustering, c-TF-IDF keyphrases, reference-grounded candidates, iterative merge at cosine ≥ 0.94.	Risk-candidate section and Algorithm 1 formalize embedding, clustering, labeling, the reference filter, and the merge rule, yielding the 182-card L4 set.
Card schema: layer-based stable IDs; bilingual fields, hierarchy, references, severity/probability; new cards need independent references, semantic distance, and a valid L2.	Generation and card-schema sections specify ID prefixes, required fields, the five-reference cap, duplicate exclusion, and release validation.
L3 mapping: with seeds, assign each L4 to the nearest L3 and check alternatives; without seeds, induce L3 families from clusters of three or more L4 risks.	L3-assignment section and Algorithm 2 formalize seeded assignment, EM-style centroid update, and the bottom-up induction fallback.
3H1R application: read Helpful, Harmless, Honest, and Role Consistency as behavior, safety-constraint, world-state, uncertainty, and role-boundary failures.	3H1R section operationalizes 3H1R, defines Primary/Secondary causal-proximity rules, and separates L4 coding from L3 icons.

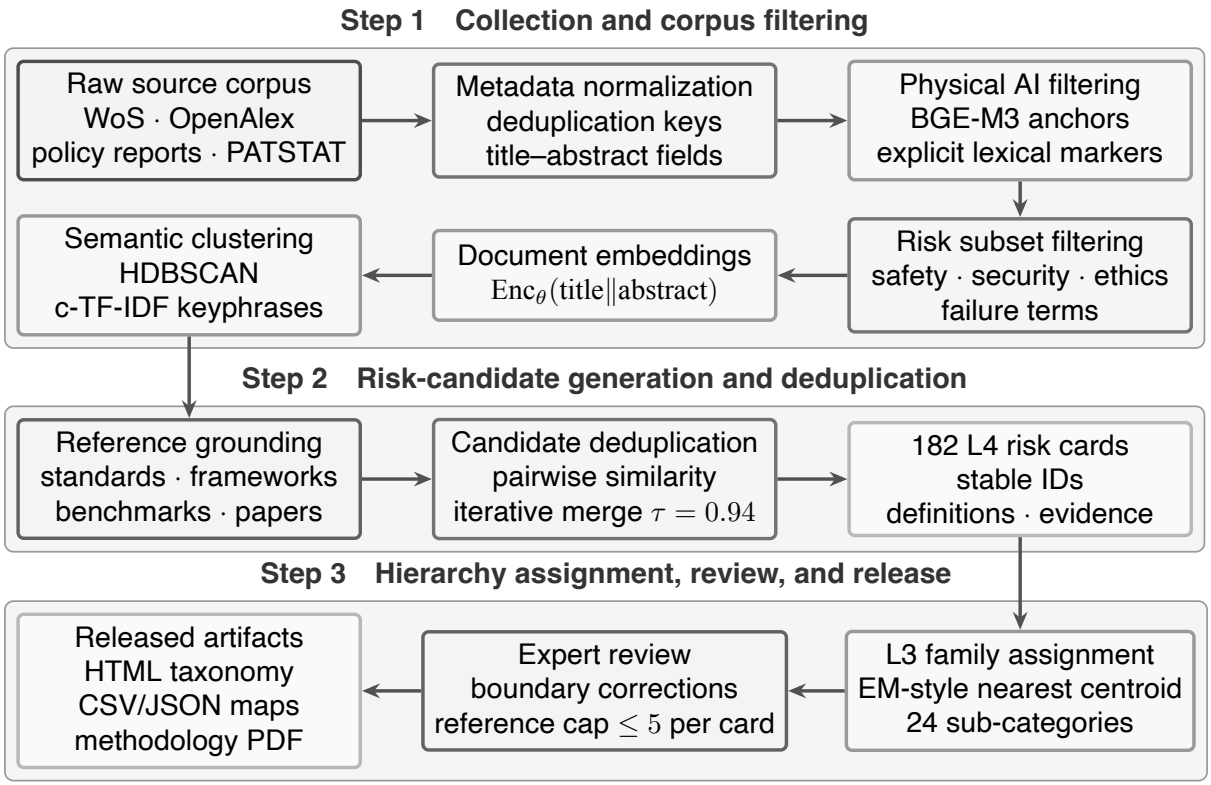

Figure 2: End-to-end taxonomy construction and release workflow

Note: Arrows denote the main data-processing path. The expert-review stage is applied iteratively: L4 risk cards and L3 assignments are repeatedly revised through expert review before release.

6 Data Sources and Corpus Construction

The construction corpus integrates three source families: scholarly publications, policy/report documents, and patent records. Scholarly records were collected from Web of Science and related local scholarly metadata snapshots, patent records from EPO PATSTAT, and policy/report records from domestic and international policy collections [44, 45, 46]. The reproducible retrieval evidence is retained in three forms, reproduced verbatim in Appendices A–C: the Web of Science topic-search expression, the PATSTAT AI patent-retrieval SQL, and the SQL-equivalent Physical AI / Physical AI risk filtering rules. Table 3 reports the integrated corpus sizes; these are the corpus-level counts used to motivate the L4 generation pipeline.

Table 3: Integrated corpus counts by source family and category

Dataset group	Papers	Policy reports	Patents	Total
AI	1,749,159	125,372	2,032,236	3,906,767
AI Infrastructure	219,002	341	522,085	741,428
Physical AI	31,386	1,654	350,109	383,149
Physical AI Risks	50,375	938	31,658	82,971

Note: Counts are reported after within-group deduplication; category overlap is intentionally retained across AI, AI infrastructure, Physical AI, and Physical AI risks. The integrated datasets are organized as three source-family collections, namely papers, policy reports, and patents, each split into AI, AI infrastructure, Physical AI, and Physical AI risks. Counts exclude records without usable title text during standardization.

Table 3 reports the integrated corpus sizes by category and source family. Paper counts are parsed from

the local Web of Science AI export, patents from the PATSTAT datasets, and policy records from the curated policy corpora, with within-group (per document type and category) deduplication and intentional cross-category overlap. Every group preserves a domestic/global key, namely applicant country for patents, Korea affiliation for papers, and a domestic/foreign flag for policy, so each figure can be reported as domestic, overseas, and total. Within the Physical AI risk group of 82,971 records, scholarly publications dominate (Web of Science, 50,375; 60.7%), followed by patents (31,658; 38.2%) and policy or report sources (938; 1.1%).

To characterize provenance beyond raw counts, country and institution were back-filled for the Physical AI risk *papers and policy reports* by re-joining each record to its original source (Web of Science / OpenAlex affiliations by DOI and source identifier, with OpenAlex and Crossref resolving the remaining gaps). This recovered an institution for 38,008 records (papers 73.6%, policy 98.7%) and a country for 36,643 records. Patents were separately enriched with applicant/assignee and country from PATSTAT (`tls207_pers_appln` $\bowtie$ `tls206_person` on `appln_id`; applicant 99.6%, country 38%), but are deliberately excluded from the country and institution rankings below: patent-applicant counts index commercial IP-filing activity and would read as a technology-competitiveness ranking rather than research and governance attention to Physical AI risks. Tables 4–5 therefore reflect scholarly papers and policy reports; counts are document–organization and document–country links over records with a resolved affiliation, not a complete census.

Table 4: Country distribution of Physical AI risk records

Rank	Country	Records	Share (%)
1	China	7,161	19.5
2	United States	6,937	18.9
3	India	2,493	6.8
4	**South Korea**	**2,208**	**6.0**
5	United Kingdom	1,697	4.6
6	Germany	1,434	3.9
7	Italy	1,121	3.1
8	Canada	1,012	2.8
9	Japan	839	2.3
10	Australia	735	2.0
11	Spain	648	1.8
12	France	537	1.5
13	Saudi Arabia	410	1.1
14	Netherlands	385	1.1
15	Pakistan	372	1.0

Note: Counts are over Physical AI risk papers and policy reports with a resolved country. Patents were enriched from PATSTAT (country present for 38% of patents) but are excluded here, because patent-applicant counts reflect commercial IP filing rather than research or governance attention and could be misread as technology competitiveness.

Physical AI risk activity is globally distributed and led by China, the United States, and India; Korea-affiliated records account for 6.0% (2,208 of 36,643 records with a resolved country).

The filtering logic has two stages. First, candidate Physical AI records are identified by cosine similarity to Physical AI definition anchors derived from NVIDIA, NIST, ISO, MITRE ATLAS, and ASIMOV-style embodied safety references[2] using the BGE-M3 encoder[3]. These anchors cover AI risk management, industrial robot safety, functional safety, controllability, adversarial AI, world-model-based Physical AI development, and embodied semantic safety [37, 48, 49, 50, 51, 38, 1, 2, 3]. Records with

[2]ASIMOV Benchmark v2: https://asimov-benchmark.github.io/v2/.

[3]BGE-M3 model card: https://huggingface.co/BAAI/bge-m3. The underlying M3-Embedding paper is cited in the References [47].

Table 5: Top institutions in Physical AI risk records

Rank	Institution / organization	Country	Records
1	Tsinghua University	China	177
2	Korea Advanced Institute of Science and Techno	South Korea	165
3	Shanghai Jiao Tong University	China	160
4	Harbin Institute of Technology	China	160
5	Technical University of Munich	Germany	142
6	Carnegie Mellon University	United States	142
7	Beijing Institute of Technology	China	141
8	Beihang University	China	133
9	Nanyang Technological University	Singapore	132
10	Seoul National University	South Korea	132
11	University of California	United States	125
12	Northeastern University	China	125
13	Chinese Academy of Sciences	China	120
14	Tongji University	China	119
15	Northwestern Polytechnical University	China	117

Note: Most frequent institutions/organizations over Physical AI risk papers and policy reports with a resolved affiliation. Patent applicants/assignees are excluded (enriched from PATSTAT as a backup) to avoid conflating commercial IP filing with research and governance attention.

similarity above 0.52, or with explicit Physical AI terms such as *physical AI*, *humanoid*, *embodied AI*, or cyber-physical systems (*CPS*), are retained as Physical AI candidates. Second, the candidate set is filtered to records whose title or abstract contains risk, safety, failure, security, ethics, or related terms. This produces the Physical AI risk group $|\mathcal{D}_{\text{PAI}}| = 82,971$. To reproduce this step, a researcher concatenates each record's title and abstract, computes the BGE-M3 embedding, compares it with the documented Physical AI anchor descriptions, applies the 0.52 threshold and lexical-marker override, and then applies the risk-term filter to the retained candidate set.

Formally, let $\mathcal{D}_{AI} = \{d_i\}_{i=1}^{N}$ be the retrieved AI corpus and let $x_i = \text{title}(d_i) \| \text{abstract}(d_i) \| \text{keywords}(d_i)$ be the normalized document text. Let $\mathcal{P} = \{p_q\}_{q=1}^{Q}$ denote Physical AI anchor descriptions derived from the authoritative sources above, and let $\mathcal{K}_{PAI}$ and $\mathcal{K}_{risk}$ denote the explicit Physical AI lexical markers and risk/safety lexical markers, respectively. The document and anchor embeddings are

$$e_i = \frac{\text{Enc}_\theta(x_i)}{\|\text{Enc}_\theta(x_i)\|_2}, \qquad u_q = \frac{\text{Enc}_\theta(p_q)}{\|\text{Enc}_\theta(p_q)\|_2}. \tag{2}$$

The Physical AI relevance score is the maximum anchor similarity:

$$s_i^{PAI} = \max_{q \in \{1,\dots,Q\}} e_i^\top u_q. \tag{3}$$

A document is retained as a Physical AI candidate if

$$d_i \in \mathcal{D}_{PAI}^{cand} \iff \left(s_i^{PAI} \geq \tau_{PAI} = 0.52\right) \vee \left[\exists k \in \mathcal{K}_{PAI} : k \subset x_i\right]. \tag{4}$$

The final Physical AI risk corpus is then

$$\mathcal{D}_{\text{PAI}} = \left\{ d_i \in \mathcal{D}_{PAI}^{cand} : \exists r \in \mathcal{K}_{risk} \text{ such that } r \subset x_i \right\}. \tag{5}$$

This two-stage rule separates semantic inclusion in the Physical AI domain from risk-specific filtering. The lexical override prevents highly specific terms such as *humanoid*, *embodied AI*, or *CPS* from being lost when a short abstract produces a lower anchor similarity score.

7 Corpus and Taxonomy Statistics

Record-level snapshot. The year-level statistics in this section are computed on a released local reproducible snapshot of 23,191 Physical AI records (7,132 scholarly, 1,042 policy, and 15,017 patent), which is the auditable slice available for direct recomputation. The full corpus totals are reported separately in Table 3 as the integrated-source view.

Level distribution. Figure 3 summarizes how the 182 released L4 risk cards distribute across the taxonomy levels. System Safety (P2) holds 92 cards (50.5%), Interaction Safety (I2) 60 (33.0%), and Societal Safety (S2) 30 (16.5%). At L3, the largest families are Robot Control (46), Accidental Harm (22), and Purposeful/Malicious Harm (17), followed by Lack of Robustness in Unseen Environments (13), Lack of Accountability & Liability (13), and Human Interaction & Safety Protocol Failures (12). Societal families are intentionally retained even when sparse because the taxonomy supports governance review and future expansion, not only frequency-based clustering.

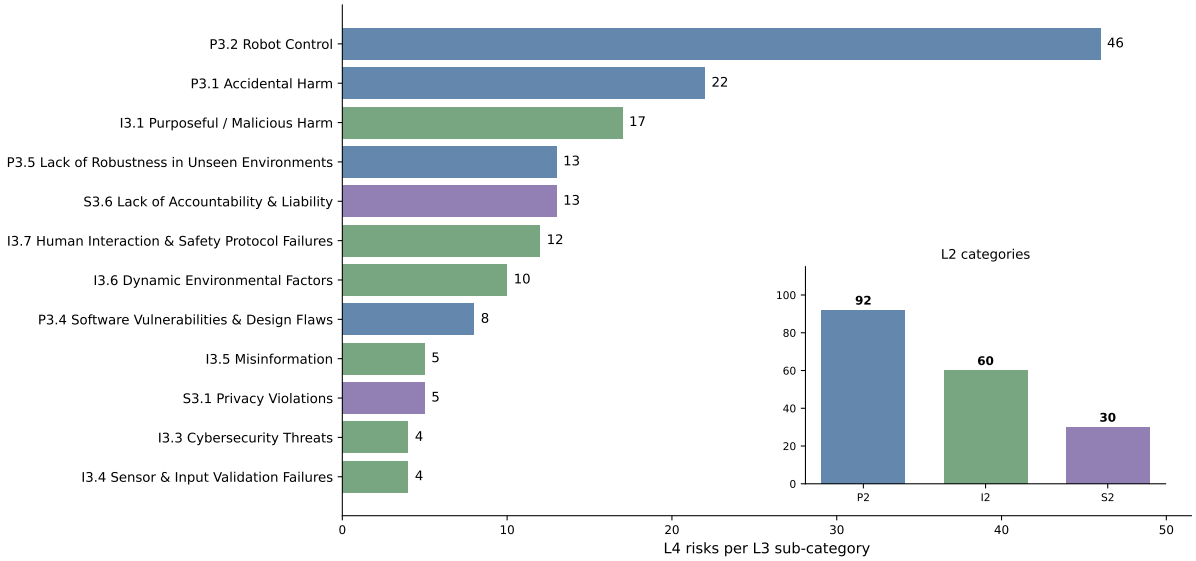

Figure 3: Distribution of the 182 released L4 risk cards

Note: Main panel: top L3 sub-categories by L4 card count, colored by L2 domain (blue = System, green = Interaction, purple = Societal). Inset: L2 category distribution (P2 System, I2 Interaction, S2 Societal).

8 Document-Space Embedding and Projection

The integrated AI document space in Figure 4 visualizes how AI, AI infrastructure, Physical AI, and Physical AI risk evidence relate at the document level. Each document d is represented by the concatenation of its title and abstract, $x_d = \text{title}(d) \| \text{abstract}(d)$, encoded with the BGE-M3 sentence embedder [47] and L_2-normalized,

$$z_d = \frac{\text{Enc}_\theta(x_d)}{\|\text{Enc}_\theta(x_d)\|_2}, \qquad \|z_d\|_2 = 1, \tag{6}$$

so that inner products $z_d^\top z_{d'}$ equal cosine similarities. Because embedding every one of the 5.1 million records in a single vector space is limited by available compute, the figure embeds a 1% stratified random sample drawn at the same rate from each source-family and category stratum of the full corpus under a fixed seed. Applying one common rate to every stratum keeps the plotted point counts proportional to the Table 3 populations, and the sample spans all years rather than a recent window.

The sampled embeddings are projected to two dimensions with uniform manifold approximation and

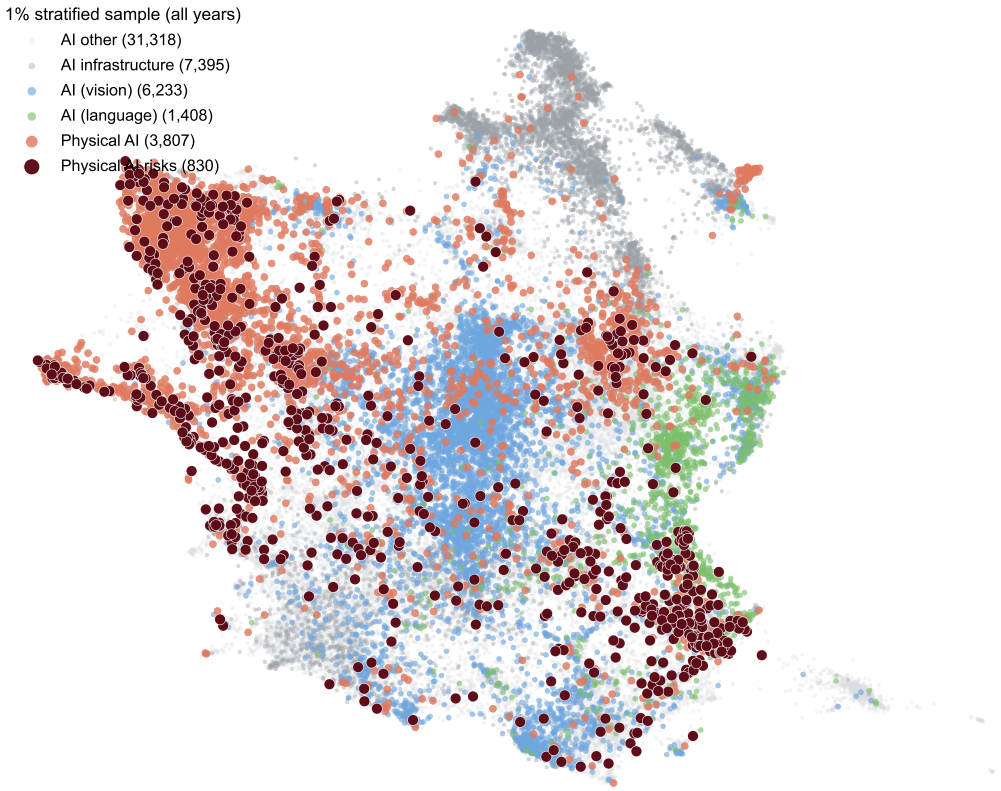

Figure 4: Physical AI document space (1% stratified sample of the full corpus, all years)

Note: The figure embeds a 1% stratified random sample of the full integrated corpus, about 50,991 of 5,099,045 records, drawn at the same rate from every source-family and category stratum so the plotted counts stay proportional to Table 3 across all years. Text is embedded with BGE-M3 (title + abstract) and projected with UMAP (fixed seed 20260704). The AI records are further split by title keywords into a vision group (6,233 points) and a broadly defined language group (1,408 points, covering natural-language processing, large language and foundation models, and generative text work), with the remaining general AI work (31,318 points) and AI infrastructure (7,395 points) forming a neutral backdrop. AI infrastructure is drawn in pure grey and the general AI work in a very light grey, while the vision and language groups carry blue and green tints so the two can be compared. Physical AI (3,807 points) and Physical AI risks (830 points) share one warm family, with Physical AI in coral and the risk layer drawn on top in a deep crimson. Read against the two coloured groups, the risk points fall predominantly in and around the vision region rather than the language region.

projection (UMAP) [52]. UMAP first builds a weighted k-nearest-neighbor graph: for point i with neighbor j, the directed membership is

$$p_{j|i} = \exp\left(-\frac{\max(0,\, d(z_i, z_j) - \rho_i)}{\sigma_i}\right), \tag{7}$$

where d is the cosine distance, ρ_i is the distance to i's nearest neighbor, and σ_i is set so that $\sum_j p_{j|i} = \log_2 k$. Directed memberships are symmetrized as $p_{ij} = p_{j|i} + p_{i|j} - p_{j|i}\, p_{i|j}$. A low-dimensional layout $\{y_i\} \subset \mathbb{R}^2$, with membership $q_{ij} = (1 + a\|y_i - y_j\|_2^{2b})^{-1}$, is optimized by minimizing the fuzzy-set cross-entropy

$$\mathcal{C} = \sum_{i \neq j} \left[p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \right], \tag{8}$$

which preserves local neighborhood structure while allowing globally distinct groups to separate. Points are colored by dataset group and marked by document type; the sample index, embeddings, and 2D coordinates are stored with the integrated dataset so that the figure is fully reproducible.

9 Risk Candidate Generation and Deduplication

The L4 generation stage transforms heterogeneous documents into semantically coherent risk candidates. Documents are embedded using BGE-M3 from title and abstract fields [47]. Hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is applied to the embedding space to identify clusters of semantically coherent risk evidence [53, 54]. Let $E = \{e_i\}_{i=1}^{|\mathcal{D}_{\text{PAI}}|}$ be the unit-normalized document embeddings for the retained Physical AI risk corpus. The base semantic distance is cosine distance:

$$\delta(i, j) = 1 - e_i^\top e_j. \tag{9}$$

For HDBSCAN, let $c_k(i)$ denote the core distance from point i to its k-th nearest neighbor. The mutual reachability distance used to build the density hierarchy is

$$d_{\text{mreach},k}(i, j) = \max\{c_k(i), c_k(j), \delta(i, j)\}. \tag{10}$$

HDBSCAN constructs a minimum spanning tree over $d_{\text{mreach},k}$, condenses the resulting hierarchy by minimum cluster size, and selects clusters with high persistence. The output is a partial assignment

$$h : \mathcal{D}_{PAI} \to \{1, \ldots, M\} \cup \{-1\}, \tag{11}$$

where $h(d_i) = m$ means that document d_i belongs to risk-evidence cluster C_m, and $h(d_i) = -1$ denotes density noise. Noise points are not discarded automatically; they can later be recovered by direct reference matching or expert review.

For each cluster C_m, class-based term frequency–inverse document frequency (c-TF-IDF) extracts representative keyphrases, following the topic-representation logic of BERTopic [55]:

$$\text{cTFIDF}(t, C_m) = \frac{f_{t,C_m}}{|C_m|} \cdot \log \frac{|\mathcal{D}_{\text{PAI}}|}{\sum_{m'} f_{t,C_{m'}}}. \tag{12}$$

The top-k keyphrases become raw risk labels. Candidate labels are retained only when they can be grounded in an authoritative reference corpus or a direct citation link. The authoritative corpus includes AI risk-management standards, robotics safety standards, adversarial-AI knowledge bases, and embodied-AI benchmarks [37, 48, 49, 50, 51, 38, 2, 3]. This reference filter reduces noisy cluster labels and ensures that each candidate is auditable. If $a_b \in \mathcal{A}$ is an authoritative reference text and ℓ_m is the raw cluster label, the semantic grounding score is

$$g_m = \max_{a_b \in \mathcal{A}} \text{sim}_{\cos}(\text{Enc}_\theta(\ell_m), \text{Enc}_\theta(a_b)). \tag{13}$$

The retained candidate set is

$$C_0 = \{c_m : g_m \geq \tau_{\text{ref}} \ \vee \ \text{directCitation}(c_m) = 1\}. \tag{14}$$

The reference-filtered candidates are then embedded again using label and definition text. For each $c_j \in C_0$, define $v_j = \text{Enc}_\theta(\text{label}(c_j) \| \text{definition}(c_j))$ and normalize v_j to unit length. Pairwise cosine similarity is

$$S_{jk} = v_j^\top v_k, \qquad j \neq k. \tag{15}$$

Deduplication is a constrained agglomerative merge over candidate nodes. At iteration t, choose

$$(j^*, k^*) = \arg\max_{j \neq k} S_{jk}. \tag{16}$$

If $S_{j^*k^*} \geq \tau_{\text{merge}} = 0.94$, the two candidates are replaced by a canonical node

$$c_{j^*} \leftarrow \text{merge}(c_{j^*}, c_{k^*}) = \big(\text{richerDefinition}, \ \text{evidence}_{j^*} \cup \text{evidence}_{k^*}, \ \text{refs}_{j^*} \cup \text{refs}_{k^*}\big), \tag{17}$$

and c_{k^*} is removed. The affected similarity row and column are recomputed from the updated label-definition text. The algorithm stops when $\max_{j \neq k} S_{jk} < 0.94$. The threshold is treated as an empirical distribution cut point: lower values collapse distinct risk types, while higher values leave near-duplicate cards.

Algorithm 1. Risk Candidate Generation and Deduplication Pipeline

1. **Input:** Tri-domain Physical AI risk corpus $\mathcal{D}_{\text{PAI}}$, with $|\mathcal{D}_{\text{PAI}}| = 82,971$ records (50,375 scholarly records, 938 policy/report records, and 31,658 patents); BGE-M3 encoder Enc_θ [47]; and an authoritative reference corpus $\mathcal{A}$ containing standards, risk frameworks, benchmarks, and high-quality papers [37, 48, 49, 38, 2].
2. **Normalize records:** Retain source identifier, title, abstract, year, DOI or URL, source family, and retrieval provenance. Remove exact duplicate records using DOI where available, otherwise title–URL keys.
3. **Embed documents:** For each $d_i \in \mathcal{D}_{\text{PAI}}$, compute $e_{d_i} \leftarrow \text{Enc}_\theta(\text{title}(d_i) \| \text{abstract}(d_i))$. Empty abstracts are represented by title-only text and flagged in the metadata.
4. **Cluster:** Apply HDBSCAN to $\{e_{d_i}\}$ and obtain semantically coherent clusters $\{C_1, \ldots, C_M\}$ [53, 54]. Treat noise points as unassigned evidence unless later recovered by direct reference matching.
5. **Extract keyphrases:** For each C_m, compute c-TF-IDF and extract the top-k terms as the raw risk label ℓ_m [55].
6. **Reference filter:** Retain ℓ_m if an authoritative source $a \in \mathcal{A}$ satisfies $\text{sim}_{\cos}(e_{\ell_m}, e_a) \geq \tau_{\text{ref}}$, or if a direct citation link exists. The retained raw candidate set is C_0, with $|C_0| = 812$. Store the matched reference, similarity score or citation link, and a short evidence sentence.
7. **Embed candidates:** For each candidate $c_j \in C_0$, compute $e_{c_j} \leftarrow \text{Enc}_\theta(\text{label}(c_j) \| \text{definition}(c_j))$.
8. **Pairwise similarity:** Compute $S_{jk} = \text{sim}_{\cos}(e_{c_j}, e_{c_k})$ for all $j \neq k$.
9. **Iterative merging:** While there exists a pair (j, k) such that $S_{jk} \geq \tau_{\text{merge}} = 0.94$, select the highest-similarity pair, merge it into a canonical node by unioning evidence and retaining the richer definition, remove the redundant node, and update the affected row and column of S.

10. **Human audit:** Review high-similarity non-merged pairs, low-reference candidates, and candidates whose evidence does not support the proposed risk definition. Remove unsupported candidates or rewrite definitions before release.
11. **Output and checks:** Export 182 unique L4 risk cards. Verify stable IDs, no duplicate DOI/title/URL references within a card, at most five references per card, and a non-empty bilingual definition and evidence justification for each released card.

10 Card Schema and Identifier Assignment

Each retained L4 risk is converted into a card with a stable identifier and a structured schema. The identifier prefix records the primary evidence layer: `PHYSRISK` for standards/frameworks, `PHYSBENCH` for benchmarks, `PHYSKR` for Korean policy evidence, and `PHYSCONN` for connectivity or telecom-infrastructure risks. The standards/framework layer is anchored in AI risk-management, robot safety, functional safety, controllability, and adversarial-AI references [37, 48, 49, 50, 51, 38]. The benchmark layer includes embodied semantic-safety and physical-safety benchmark work [2, 3]. The current release contains 55 `PHYSRISK`, 113 `PHYSBENCH`, 5 `PHYSKR`, and 9 `PHYSCONN` cards.

The required card fields are: L4 identifier, Korean and English label, Korean and English risk definition, L2/L3 hierarchy, severity proxy, probability proxy, evidence justification, reference title/link, and 3H1R alignment tags. The 3H1R tags extend helpful-honest-harmless alignment concepts and constitutional and role consistency ideas to Physical AI settings [14, 15, 2]. Reference curation follows a maximum of five references per card, with duplicate DOI/title/URL exclusion and card-specific evidence justifications.

New-card admission follows a stricter governance rule for final taxonomy expansion. A newly proposed L4 card is admitted only when three conditions are jointly satisfied: (i) it has at least three independent references or evidence sources after duplicate DOI/title/URL removal; (ii) its semantic similarity to existing cards remains below the merge threshold $\tau_{\text{merge}} = 0.94$; and (iii) it can be assigned to a valid L2 category without violating the L2 responsibility-boundary definition. This rule separates initial candidate discovery, where one authoritative reference is sufficient to retain a raw candidate for review, from final taxonomy expansion, where stronger independent support is required.

11 Hierarchical Assignment to L3 Families

The L3 hierarchy can be assigned in two modes. In the current release, 24 L3 seed descriptions are defined before L4 placement. For each new L4 card, the label and definition are embedded and compared against L3 anchors; the card is assigned to the nearest L3 family, with expert review for close or semantically ambiguous cases. If no L3 seed hierarchy exists, L4 cards can be clustered first and candidate L3 families can be induced when three or more coherent L4 cards form a stable family. This reflects the same hybrid principle used in prior AI-risk taxonomy synthesis, where expert taxonomic structure and evidence-derived candidate groups are reconciled rather than treated as substitutes [5, 6].

Algorithm 2 formalizes the iterative assignment process as an expectation–maximization (EM)-style procedure [56, 57].[4] Let $R = \{r_i\}_{i=1}^{182}$ denote the released L4 risks and let $K = 24$ be the number of L3

[4]The assignment is the classical Expectation–Maximization algorithm [56] specialized to unit-norm (cosine) document representations, i.e. spherical k-means / concept decomposition [57]: the E-step assigns each L4 card to its nearest L3 centroid, and the M-step recomputes and renormalizes each centroid to the unit sphere.

Figure 5: Example L4 risk-card rendering in the released taxonomy interface.

Note: The screenshot shows how a released L4 card combines a stable identifier, bilingual risk label and definition, evidence justification, references, severity and probability proxies, and 3H1R Primary/Secondary tags. The image documents the rendered card schema in the release interface.

seed families. Each L4 risk is represented by a unit vector

$$y_i = \frac{\text{Enc}_\theta(\text{label}(r_i) \| \text{definition}(r_i) \| \text{evidenceKeywords}(r_i))}{\|\text{Enc}_\theta(\text{label}(r_i) \| \text{definition}(r_i) \| \text{evidenceKeywords}(r_i))\|_2}. \tag{18}$$

Each L3 family is represented by a unit centroid μ_k. The latent assignment variable $z_i \in \{1, \ldots, K\}$ indicates the L3 family of r_i. The assignment objective is spherical k-means style maximization of within-family cosine similarity:

$$\max_{\{z_i\},\{\mu_k\}} \sum_{i=1}^{182} y_i^\top \mu_{z_i} \quad \text{s.t.} \quad \|\mu_k\|_2 = 1 \ \ \forall k. \tag{19}$$

The E-step assigns each risk to the nearest L3 centroid in cosine space:

$$z_i^{(t+1)} = \arg\max_{k \in \{1,\ldots,K\}} y_i^\top \mu_k^{(t)}. \tag{20}$$

The M-step updates each centroid as the normalized mean of assigned L4 vectors:

$$\mu_k^{(t+1)} = \frac{\sum_{i: z_i^{(t+1)} = k} y_i}{\left\| \sum_{i: z_i^{(t+1)} = k} y_i \right\|_2}, \tag{21}$$

with the seed centroid retained when a family receives no assigned L4 risk. Expert review is retained as a final correction layer because some Physical AI risks are semantically close but operationally distinct. This human correction step is a governance constraint that prevents semantically plausible but operationally invalid assignments.

Algorithm 2. EM-Based L3 Family Assignment

1. **Input:** L4 embeddings $\{e_{r_i}\}_{i=1}^{182}$ and L3 seed descriptions $\{d_k\}_{k=1}^{K}$, with $K = 24$ in the current release.
2. **Prepare L4 text:** For each risk r_i, concatenate the bilingual label, compressed English definition, Korean definition, and evidence keywords. Normalize whitespace and keep the stable L4 identifier as the row key.
3. **Initialize:** Set $\mu_k^{(0)} \leftarrow \text{Enc}_\theta(d_k)$ for $k = 1, \ldots, K$, normalize centroids to the unit sphere, and set $t \leftarrow 0$.
4. **E-step:** Assign $z_i^{(t+1)} \leftarrow \arg\max_k \text{sim}_{\cos}(e_{r_i}, \mu_k^{(t)})$ for all L4 cards i.
5. **M-step:** Update each centroid $\mu_k^{(t+1)}$ as the normalized mean of the embeddings assigned to family k. If a family becomes empty, retain its seed centroid and flag the family for expert review.
6. **Convergence check:** If $z^{(t+1)} = z^{(t)}$, stop; otherwise set $t \leftarrow t + 1$ and repeat the E-step.
7. **Expert review:** Human reviewers inspect assignments deviating from domain expectations, especially cards with close first- and second-choice L3 similarities, sparse L3 families, or L3 definitions that no longer match their assigned L4 risks.
8. **Coverage constraint:** Ensure every L3 family retains at least one L4 risk. When a family is empty, reassign the nearest semantically valid L4 risk only after checking that the L3 definition and L4 risk definition are substantively compatible.
9. **Output and checks:** Export final L3 assignments $\{z_i\}_{i=1}^{182}$, update L2/L3 counts, and verify that

the released HTML, CSV, JSON, and reference-map artifacts report the same hierarchy.

12 Released Hierarchy

The released hierarchy has three L2 categories and 24 L3 sub-categories. Table 6 summarizes the L2 distribution.

Table 6: L2 distribution of L4 risk cards

L2 category	L4 cards	Share
P2 System Safety	92	50.5%
I2 Interaction Safety	60	33.0%
S2 Societal Safety	30	16.5%
Total	182	100.0%

The largest L3 family is *Robot Control* with 45 cards, followed by *Accidental Harm* with 22 cards and *Purposeful/Malicious Harm* with 17 cards. Societal categories are intentionally retained even when sparse because the taxonomy is designed to support governance review and future expansion, not only frequency-based clustering.

13 Validation and Quality Controls

Quality controls operate at three levels. At the corpus level, the pipeline filters candidate records through Physical AI anchors, explicit lexical markers, and risk/safety/security terms. At the candidate level, semantic duplicates are merged with a conservative threshold, and candidate cards require authoritative references. At the released-card level, the export script verifies card counts, L3/L2 distributions, reference rows, linked references, maximum reference count per card, and evidence/reference mismatches. These controls are intended to preserve the traceability emphasized by AI risk-management frameworks and structured AI risk repositories [37, 48, 5].

The final HTML export currently reports 182 L4 cards, 360 evidence/reference rows, and no card exceeding the five-reference cap. Eight cards contain a mismatch between the number of evidence clauses and the number of linked reference rows (a distinct metric from the seven unlinked evidence rows in Table 15). These mismatches are reported rather than corrected, since they reflect the current HTML state.

14 Sensitivity Analysis

We test whether the two data-driven stages of the pipeline, density-based candidate clustering and the L3 family assignment, remain stable under changes to their hyperparameters and under perturbation of the inputs.

Embedding and clustering hyperparameters

The candidate-generation pipeline is governed by a small, fixed set of hyperparameters, all recorded with a single random seed (20260704) for reproducibility. The BGE-M3 encoder [47] is run in dense mode,

producing a 1024-dimensional vector per document; inputs are the concatenated title and abstract truncated to 1,200 characters, encoded with a maximum sequence length of 256 tokens, batch size 32, and L_2 normalization so that inner products equal cosine similarities. For density-based clustering the embeddings are first reduced with UMAP [52] configured as n_neighbors $= 30$, min_dist $= 0.0$, ten output components, and cosine metric, a standard configuration for downstream clustering. HDBSCAN [53, 54] is then applied with min_cluster_size $= 15$, min_samples $= 5$, Euclidean metric on the reduced space (equivalent to cosine on the normalized inputs), and excess-of-mass (`eom`) cluster selection. Downstream thresholds are the reference-grounding cut τ_{ref} and the deduplication merge threshold $\tau_{\text{merge}} = 0.94$.

Clustering robustness

Because density-based clustering can be sensitive to its settings, the configuration was stress-tested on a reproducible sample of 4,000 Physical AI risk documents (drawn across papers, policy reports, and patents under the fixed seed). The two HDBSCAN knobs were swept over min_cluster_size $\in \{8, 10, 15, 20, 30\}$ and min_samples $\in \{1, 5, 10\}$, and the UMAP neighborhood over n_neighbors $\in \{15, 30, 50, 80\}$; each configuration is compared to the baseline partition by the Adjusted Rand Index (ARI). Table 7 reports the results. The macro-structure is stable for the neighborhood of settings around the baseline. Holding min_samples $= 5$, varying min_cluster_size from 8 to 20 keeps ARI in the 0.86–0.95 range, and the cluster count moves smoothly (58 to 36) without fragmenting or collapsing. Sensitivity is concentrated in min_samples $= 1$ (which admits many small, low-density clusters, ARI ≈ 0.43–0.55) and in large UMAP neighborhoods that over-smooth local density. Because this stage only proposes L4 *candidates*, which are later filtered against the authoritative reference corpus, deduplicated at τ_{merge}, and confirmed by human audit, the released 182-card taxonomy does not depend on any single clustering resolution.

Convergence and stability of the L3 assignment

The EM-style assignment of Algorithm 2 was validated by a convergence and stability simulation on the released 182 L4 cards, embedded from their label and definition text with BGE-M3 under a fixed seed. Three checks are reported in Figure 6, and they ask, in turn, whether the procedure terminates, whether its output agrees with the geometry of the embedding space, and whether that output survives small perturbations of the input. First, the spherical k-means and expectation-maximization iteration reaches a fixed point in seven steps, with the within-family cosine objective increasing monotonically and the number of reassigned cards falling to zero, the termination behaviour that alternating-optimization assignment of this kind is expected to show [57, 56]. Second, the released partition is internally consistent with the embedding space, which we measure through top-k containment, the fraction of cards whose released L3 family lies among their k nearest family centroids. The released L3 is the single nearest family for 89.6% of cards, within the top two for 98.9%, within the top three for 99.5%, and within the top five for all cards, so the assignment reaches the 90%, 95%, and 97% levels once a one-to-two-rank tolerance is allowed, consistent with the expert-review step reconciling the small fraction of adjacent-family cases. A label-permutation test corroborates that the grouping is not an artefact of chance, since the observed mean within-family cohesion of 0.811, a silhouette-style summary of how tightly cards sit within their assigned family [58], exceeds the cohesion of random partitions of identical family sizes (0.788) with $p < 0.001$ over 5,000 permutations. Third, the assignment is robust to embedding perturbation, following the logic of stability-based cluster validation, under which a trustworthy partition should persist when the data are slightly perturbed [59]. Adding isotropic Gaussian noise of standard deviation σ to the card embeddings and reassigning, agreement with the unperturbed assignment stays at or above 97% up to $\sigma \approx 0.05$ and above 90% up to $\sigma \approx 0.075$, a noise scale that already exceeds the representational jitter expected from paraphrase-level text changes.

Post-audit reassignment sensitivity. A subsequent full wording and boundary audit revised 84 L4 labels

Table 7: HDBSCAN and UMAP hyperparameter robustness (4,000-document risk sample)

Configuration	Clusters	Noise	ARI vs base
HDBSCAN sweep (UMAP n_neighbors=30)			
mcs=8, ms=1	109	0.28	0.43
mcs=8, ms=5	58	0.30	0.86
mcs=8, ms=10	46	0.39	0.65
mcs=10, ms=1	85	0.28	0.47
mcs=10, ms=5	54	0.32	0.85
mcs=10, ms=10	44	0.39	0.65
mcs=15, ms=1	48	0.22	0.54
mcs=15, ms=5	41	0.31	1.00
mcs=15, ms=10	38	0.37	0.70
mcs=20, ms=1	42	0.22	0.55
mcs=20, ms=5	36	0.32	0.95
mcs=20, ms=10	33	0.36	0.71
mcs=30, ms=1	31	0.24	0.55
mcs=30, ms=5	27	0.27	0.69
mcs=30, ms=10	25	0.37	0.63
UMAP sweep (HDBSCAN mcs=15, ms=5)			
n_neighbors=15	44	0.27	0.44
n_neighbors=30	41	0.31	1.00
n_neighbors=50	34	0.35	0.55
n_neighbors=80	34	0.37	0.52

Note: Each row perturbs one knob relative to the baseline (UMAP $n\_neighbors = 30$; HDBSCAN min_cluster_size $= 15$, min_samples $= 5$), which by definition has ARI = 1.00. Clusters excludes the noise class; Noise is the fraction of documents left unassigned; ARI is the Adjusted Rand Index against the baseline partition on the identical 4,000-document sample. Higher ARI means more agreement with the baseline cluster structure. Here mcs and ms abbreviate the HDBSCAN hyperparameters min_cluster_size and min_samples.

or definitions and tested five candidate moves with constrained spherical EM. The sensitivity design crossed three locally available embedding models (BGE-M3, mxbai-embed-large, and nomic-embed-text), bilingual and English-only card text, and L3 seed weights $w \in \{1, 3, 5, 10\}$, for 24 conditions. Each candidate comparison used 1,000 bootstrap replicates, and every admissible assignment was required to retain at least one L4 risk in each of the 24 L3 families. A move was released only when the same target was selected in at least 80% of the sensitivity conditions. This rule moved PHYSBENCH-REF-0065 from I3.7 to P3.2 (20/24 conditions) and PHYSBENCH-REF-0107 from I3.6 to S3.6 (23/24). The other three candidates retained their released assignments. Holding revised wording fixed, the two moves increased mean micro assignment margin from 0.025601 to 0.026731 (+4.41%), reduced the negative-margin fraction from 0.246566 to 0.234203 (-5.01%), and increased leave-one-out micro cohesion from 0.780581 to 0.781258. The absolute cohesion change is small because only two of 182 risks moved. These statistics assess internal semantic consistency rather than external validity.

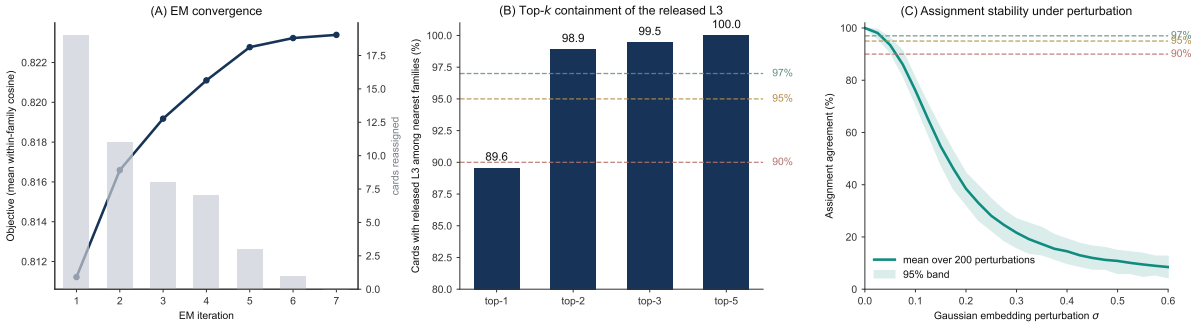

Figure 6: Convergence and stability of the 182-card L3 assignment

Note: (A) The EM iteration converges in seven steps: the within-family mean cosine (objective) rises monotonically while the number of reassigned cards (bars) drops to zero. (B) Top-k containment: fraction of cards whose released L3 family is among the k nearest family centroids (top-1 89.6%, top-2 98.9%, top-3 99.5%, top-5 100%); dashed lines mark the 90/95/97% levels. (C) Stability under Gaussian embedding perturbation of standard deviation σ (mean and 95% band over 200 perturbations per σ); agreement with the unperturbed assignment stays $\geq$97% to $\sigma \approx 0.05$. All panels use BGE-M3 embeddings of the released 182 cards under fixed seed 20260704.

15 3H1R Alignment Framework and Taxonomy Design Principles

The taxonomy uses 3H1R as an alignment-diagnosis framework for identifying which alignment value fails in a Physical AI risk. The framework has two roles. First, it defines a value axis for coding each risk in terms of Harmless, Helpful, Honest, and Role Consistency failures. Second, it provides design rules for accumulating those judgments into a reproducible risk taxonomy. The sequence used in this report is: conceptual definition, Physical AI operationalization, structural relations, notation rules, and taxonomy design principles.

15.1 Conceptual Definitions

The 3H dimensions are adapted from the helpful–honest–harmless framing of language-model alignment [14]. Role Consistency is added from constitutional-AI and embodied-agent constraint work, where role, authority, and behavioral principles must remain stable under changing contexts [15, 2, 3]. In this report, the four dimensions carry the following conceptual definitions. Harmless (H1) means that system behavior or output should not cause harm to individuals, groups, or society, covering both direct harm and indirect harm such as enabling a dangerous action, and it stands in structural tension with Helpful. Helpful (H2) means assistance aligned with the user's real intent and context, judged by practical benefit rather than mere instruction following, so that excessive refusal also counts as a helpfulness fail-

ure. Honest (H3) means an accurate representation of facts, uncertainty, and intent without deception or manipulation, which in Physical AI centers on whether the system truthfully reports its perception, uncertainty, capability, and state. Role Consistency (RC) means preserving the system's defined role, authority boundary, and behavioral principles across changing contexts, with persona drift, context violation, jailbreak acceptance, and role-boundary failure as the core failure modes. Table 8 then translates each dimension into observable Physical AI failure patterns.

15.2 Operationalization for Physical AI

Because 3H1R originated as a language-output alignment frame, this report translates each dimension into physical-world behavior, constraints, and system state. The diagnostic question asks how the value should be interpreted in physical space, while the operational definition states what observable failure pattern is coded.

Table 8: 3H1R operationalization for Physical AI

Dimension	Original LLM definition	Diagnostic question for Physical AI	Operational definition in Physical AI
H1 Harmless	Avoid harmful content generation	Can this risk directly cause physical harm?	Physical injury, property damage, environmental harm, or violation of explicit safety constraints such as speed, force, or separation distance.
H2 Helpful	Provide information aligned with user intent	Does this risk reflect a trade-off failure between task capability and safety requirements?	Failure to balance task completion and safety requirements, including over-sacrificing capability or over-prioritizing completion while ignoring safety.
H3 Honest	Truthfulness and uncertainty awareness	Does the system misperceive the world, understate uncertainty, or misrepresent its own capability or state?	Misrecognition of world state, under-reporting or under-estimation of uncertainty, or over-representation of system capability and state.
RC Role Consistency	Preserve role and value principles	Does the system fail to preserve role, authority, and safety principles under adversarial inputs or edge cases?	Departure from role or authority boundaries, jailbreak acceptance, or collapse of operational protocol and constitutional constraints.

Harmless and Role Consistency are separated analytically. RC refers to the violation of normative or authority boundaries itself, such as deviation from an operating protocol. Harmless refers to the physical harm that may follow from that violation. When both appear in the same causal chain, the Primary/Secondary distinction is used to express causal proximity.

Primary and Secondary tags are assigned by causal proximity rather than by outcome severity. If a risk definition describes a dimension as the direct mechanism of failure, the dimension is Primary. If it becomes involved as the failure unfolds, the dimension is Secondary. The default rule is one Primary tag. Two Primary tags are allowed for explicit tradeoffs or causal branching. Three Primary tags are permitted only as documented exceptions.

15.3 Structural Relations

3H1R is not a list of independent labels. It is a two-axis mapping: the taxonomy axis identifies what type of risk failed, while the alignment-value axis identifies which value was violated. A risk occupies one location in the taxonomy axis but may carry multiple non-exclusive alignment-value tags.

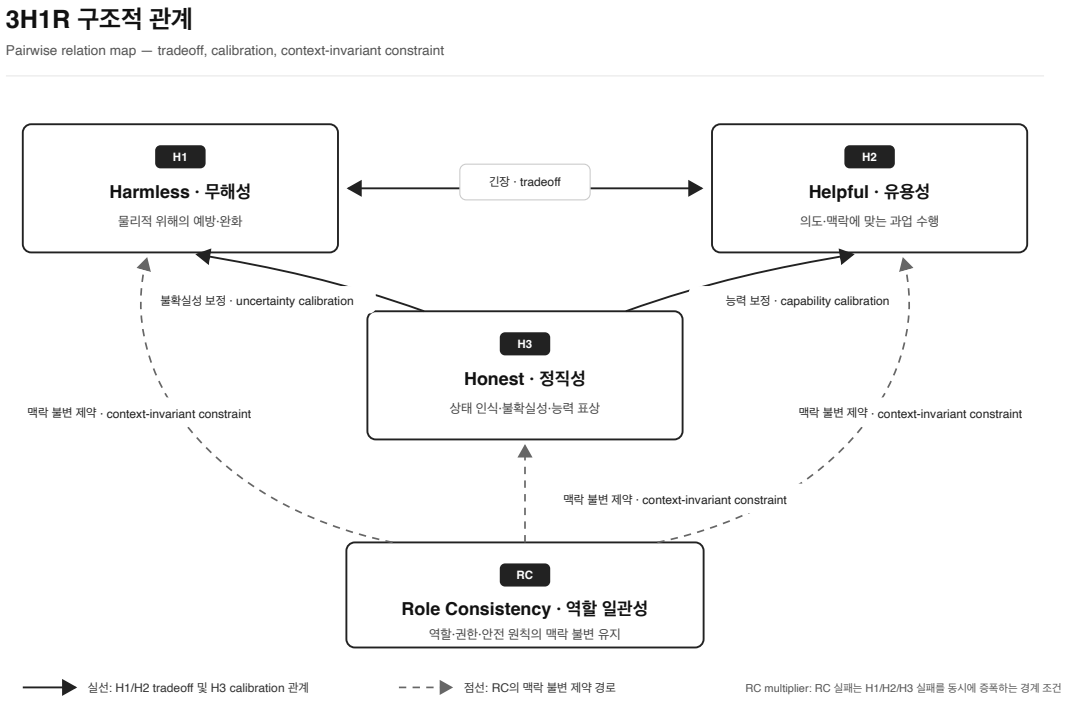

Figure 7: Pairwise structural relations among H1 Harmless, H2 Helpful, H3 Honest, and Role Consistency.

Note: Solid paths represent the H1–H2 tradeoff and H3 calibration relations. Dashed paths represent Role Consistency as a context-invariant constraint across role, authority, and safety principles. The multiplier effect of RC failure is treated as a boundary condition rather than as a separate taxonomy class.

Table 9: Structural relations among 3H1R dimensions

Relation	Scholarly term	Coding implication
H1 Harmless ↔ H2 Helpful	Tradeoff / safety–utility tension	Strengthening safety can increase over-refusal, while pushing usefulness can increase safety-constraint violation.
H3 Honest → H1 Harmless	Uncertainty calibration	Greater uncertainty about world state, sensor reliability, or self-capability should make safety judgment more conservative.
H3 Honest → H2 Helpful	Capability calibration	The system must represent what it can and cannot do in order to tune task intensity and refusal level.
RC Role → H1/H2/H3	Context-invariant constraint	Role, authority, and safety principles must not be relaxed or omitted when prompts, users, environments, or task contexts change.
RC × H1/H2/H3	Multiplier / constraint-failure amplification	If RC collapses, the admissible range of H1, H2, and H3 judgments shifts simultaneously, amplifying other alignment failures.

For example, if a robot knocks over an equipment rack, the risk is not coded solely as Harmless by observing the outcome. The mechanism is decomposed: perceptual error is an Honest failure, failure of braking or avoidance is a Harmless failure, and departure from an operating constraint is a Role Consistency failure. The Primary tag is assigned to the mechanism most causally proximate to the risk definition; the remaining dimensions are Secondary when they participate downstream.

15.4 Notation Rules Across Taxonomy Levels

3H1R notation has different purposes at different levels of the taxonomy. L4 risk cards require precise mechanism-level judgment and therefore show both Primary and Secondary tags. L3 sub-categories summarize which Primary alignment failure is most representative among their lower-level risks. L2 categories do not display 3H1R representative icons; their alignment profile is read from the L3 rows below them.

Table 10: Notation rules across taxonomy levels

Layer	Purpose	Notation rule
L4 Risk Cards	Code the direct cause and derivative alignment dimensions of a specific risk.	Display axis and grade, e.g., H1 Harmless[P] and H3 Honest[S].
L3 Sub-categories	Summarize which Primary alignment failures are most representative of the L4 risks in that sub-category.	Display one or two representative Primary icons only. Do not show Secondary, counts, or ratios.
L2 Categories	Show domain-level category and number of lower L3 sub-categories.	Do not show 3H1R representative icons at L2.

For L4 notation, dimensions are always displayed in the fixed order H1 Harmless, H2 Helpful, H3 Honest, and RC Role Consistency. This order is not a priority ranking; priority is expressed only by the Primary/Secondary grade. For L3 representative icons, if the most frequent Primary axis accounts for at least 60% of the L4 risks, a single representative icon is used. If the leading axis is below 60% and the second axis is at least 25%, two icons are shown. If three or more axes each account for at least 20%, the L3 is marked as Mixed.

15.5 Taxonomy Design Principles

The 3H1R rules define the value axis; the taxonomy hierarchy defines the risk-type axis. The design combines both axes. Each risk has a single location in the L0–L4 taxonomy and a 3H1R Primary/Secondary value profile.

Table 11: Taxonomy design principles

Principle	Implementation
Hierarchy	The taxonomy is organized as L0 Ownership Levels, L1 Physical AI Risks, L2 Categories, L3 Sub-categories, and L4 Risk Cards.
Categorization	The taxonomy aims for practical coverage and mutual exclusivity. Boundary cases are documented, with reassignment decisions and history recorded.
Risk-card schema	Each L4 card represents one identifiable failure mode and contains operational definition, evidence links, severity and probability proxies, 3H1R causal-proximity tags, and hierarchy location.
Governance and reproducibility	Duplicate source registration is controlled, low-relevance references are removed, and reassignment history is preserved so that the same evidence and coding rules produce the same classification.

16 Harmless–Helpful Trade-off for Physical AI Actions

For a physically embodied agent, the 3H1R *Harmless* and *Helpful* dimensions are coupled: acting faster and with less deliberation raises task efficiency but erodes the caution margin that keeps an action harmless [15]. We cast this as a two-axis operating space for humanoid and other Physical AI actions (Figure 8), using established formalisms from quantitative risk analysis, embodied-agent evaluation, and safe control for each axis.

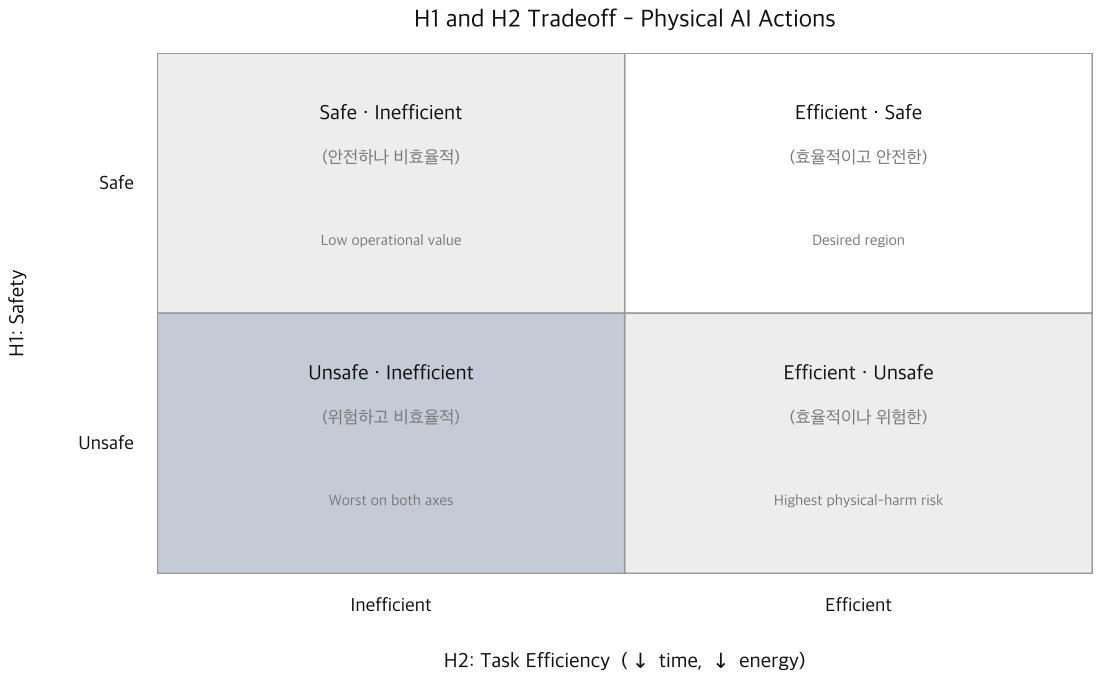

Figure 8: Harmless–helpful trade-off for Physical AI actions

Note: The vertical axis is harmlessness (safety); the horizontal axis is task efficiency, i.e. helpfulness operationalized as low time and energy *given task completion*. Shading encodes desirability rather than color semantics: white is the desired *Safe · Efficient* region, the two light-gray cells are single-axis compromises, and the darker *Unsafe · Inefficient* cell is worst on both axes. The *Unsafe · Efficient* cell is the alignment-critical one: its danger comes from capability, not incompetence.

Axis 1: harmlessness

We quantify harm with the standard set-of-triplets definition of risk, in which risk is the expected consequence aggregated over hazard scenarios [60]. For an action $a \in \mathcal{A}$ executing task τ with hazard events $\{e_k\}$ of severity $s(e_k) \geq 0$,

$$\mathrm{Harm}(a) = \mathbb{E}\left[\sum_k s(e_k)\,\mathbf{1}\{e_k \mid a\}\right], \tag{22}$$

and, following the hazardous-event severity framing of the automotive safety-of-the-intended-functionality standard [19], the safety margin against an admissible-harm budget $h_{\max}$ is $g(a) = h_{\max} - \mathrm{Harm}(a)$; a is *Safe* when $g(a) \geq 0$. This is the vertical axis of Figure 8.

Axis 2: task efficiency as helpfulness

Helpfulness for a physical agent is how economically a *completed* task consumes time and energy, which we read through two established Physical-AI measures. Energetic economy of locomotion is the dimen-

sionless cost of transport [61],

$$\mathrm{CoT} = \frac{E}{m\,g\,d}, \tag{23}$$

the energy E per unit weight mg and travelled distance d; lower is better. Task-and-navigation performance is measured by success-weighted path length [62],

$$\mathrm{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \,\frac{\ell_i^{\star}}{\max(\ell_i, \ell_i^{\star})}, \tag{24}$$

where S_i indicates success on episode i and $\ell_i, \ell_i^{\star}$ are the achieved and shortest path lengths; the success gate S_i removes the degenerate "do nothing" optimum and can be extended from path length to time or energy. Efficiency, measured as low CoT and high SPL and optionally normalized to a skilled-human baseline for anthropomorphic platforms, is the horizontal axis of Figure 8. Thresholding the two axes gives the figure's four quadrants; the alignment-critical cell is *Unsafe · Efficient*, where a highly capable action attains the task while breaching the harm budget.

Trade-off as constrained optimization

A pure efficiency objective drives behavior toward the *Unsafe · Efficient* cell, because relaxing caution lowers cost. Alignment therefore treats harmlessness as a *constraint* rather than a competing reward, the constrained-Markov-decision-process view [23] that underlies safe reinforcement learning for physically interacting systems [21]:

$$\pi^{\star} = \arg\max_{\pi}\ \mathbb{E}_{\pi}\Big[\textstyle\sum_t r_t^{\mathrm{eff}}\Big] \quad \text{s.t.} \quad \mathbb{E}_{\pi}\Big[\textstyle\sum_t \mathrm{Harm}_t\Big] \le d, \tag{25}$$

where r^{eff} rewards efficiency and cumulative harm is capped by the budget d. The unconstrained efficiency optimum may lie in the unsafe region; enforcing the harm constraint projects $\pi^{\star}$ onto the feasible safe set and lands the constrained optimum in the *Safe · Efficient* cell [21].

This is where the taxonomy re-enters. The two Unsafe rows of Figure 8 are populated by the released L4 risk cards, and each card's 3H1R tags identify which alignment value the corresponding failure violates. The matrix explains why an efficiency-shaped training or operational objective, the default for humanoid platforms, systematically pressures behavior toward the highest-harm quadrant unless an explicit harmlessness constraint is enforced. The L4 catalogue is the evidence base from which that constraint (the admissible-harm set) is instantiated.

17 Reproducibility Artifacts

The public taxonomy is maintained as a single HTML page with a versioned copy, and is exported by a release script into machine-readable card, reference, and summary files together with the curation and audit scripts and the documented 3H1R Primary-axis exception list. A reproducible rerun follows this order. First, verify the source-retrieval artifacts reproduced in the appendices: the PATSTAT SQL and the Web of Science query. Second, reconstruct or verify the local working snapshot and confirm its source-family and category denominators against Table 3. Third, apply Algorithm 1 to generate and deduplicate L4 candidates, preserving all source identifiers and reference links. Fourth, apply Algorithm 2 to assign the resulting L4 risks to the 24 L3 families and record any human corrections as audit metadata. Fifth, export the final HTML into the machine-readable card, reference, and summary files. Finally, compare the exported counts against Table 15 and verify that no L4 card violates the five-reference cap or loses its bilingual definition, evidence justification, or hierarchy assignment.

18 Limitations and Open Problems

This release is a draft, and several of its commitments are motivated rather than yet demonstrated. We state the main limitations plainly.

1. **The added Role Consistency axis is not yet empirically validated.** RC is theoretically motivated as the embodied analogue of obedience and authority, but we have not shown that annotators apply it reliably or that it captures variance not already covered by Harmlessness. Whether RC is a distinct axis or a correlated sub-case of H1 is an open question requiring a labeling study with inter-annotator agreement measured on RC specifically.
2. **3H1R coverage is asserted, not measured.** We claim the four axes span the alignment-relevant failure surface for embodied action, but offer no completeness argument; some cards, emergent multi-agent or environmental-structural risks in particular, sit awkwardly on a per-action value scheme [22].
3. **Severity and probability are proxies, not measurements.** The per-card estimates are expert-assigned ordinal judgments in the quantitative-risk tradition [60] without incident-frequency data behind them. They should be read as a triage ranking, not calibrated risk, and are the weakest input to any downstream prioritization.
4. **Corpus and source bias.** The evidence base draws on published papers in Web of Science, policy reports, and PATSTAT patents, retrieved with multilingual queries rather than English alone. The residual bias is therefore one of source type rather than language. Failure modes that live mainly in operational logs, incident databases, or gray literature remain under-represented relative to peer-reviewed and patented work, and the patent channel over-weights commercially patentable failures.
5. **The expert L3 partition is only partially reproducible from embeddings.** Top-1 embedding assignment recovers about three-quarters of the human L3 structure (see the Sensitivity Analysis); the remaining quarter depends on expert judgment the pipeline does not reconstruct, which bounds how "bottom-up" the hierarchy can honestly be called. The current partition was defined by internal domain experts, and independent consultation with external experts is still needed to confirm the L3 families and their boundaries.
6. **The 182 cards are a snapshot, not an exhaustive enumeration.** The catalogue reflects the corpus and expert pool at one point in time. Physical AI capability moves faster than the taxonomy can be re-mined, and we have not yet committed to an update cadence or versioning discipline.
7. **The Harmless–Helpful account is a conceptual model, not a fitted controller.** We argue *why* efficiency pressure produces the high-harm quadrant, but do not fit a constrained-MDP controller to embodied data or estimate the multipliers empirically; the claim is a directional, theory-level prediction [23, 21].
8. **Limited annotation validation.** Card definitions and 3H1R tags rest on a small annotator pool with limited inter-rater reliability reporting; the multi-rater study needed to separate genuine signal from individual judgment, especially for the new RC axis and borderline severity calls, remains to be done.

19 Conclusion

The Physical AI Risk Taxonomy construction process uses a hybrid method: authoritative upper-level structure, corpus-driven L4 candidate discovery, semantic deduplication, EM-style hierarchy assignment, and human review. The resulting release provides a 182-card L4 risk taxonomy that is both empirically grounded in a large multi-source corpus and operationally usable as a governance, evaluation, and benchmarking reference for Physical AI systems.

A PATSTAT AI Patent Retrieval SQL

Listing 1 reproduces the PATSTAT SQL used to retrieve the AI patent corpus before applying the Physical AI and risk filters. The listing below is imported verbatim from the preserved source query so that the SQL text remains unchanged.

```
SELECT DISTINCT
 a.appln_id app_id,
 a.appln_auth patent_office,
 a.appln_nr app_num,
 a.earliest_filing_year priority_year,
 t.appln_title title,
 b0.appln_abstract abstract,
 a.docdb_family_size family_size,
 a.granted granted,
 CAST(NULL AS INTEGER) ipc_count,
 CAST(NULL AS INTEGER) cpc_count,
 CAST(NULL AS VARCHAR(4000)) applicant_info,
 CAST(NULL AS VARCHAR(4000)) applicant_std_name_ids,
 CAST(NULL AS VARCHAR(4000)) applicant_sectors,
 CAST(NULL AS VARCHAR(4000)) applicant_ctry_codes,
 CAST(NULL AS VARCHAR(4000)) inventor_info,
 CAST(NULL AS VARCHAR(4000)) inventor_sectors,
 CAST(NULL AS VARCHAR(4000)) inventor_ctry_codes,
 CAST(NULL AS VARCHAR(4000)) all_ipc_codes,
 CASE WHEN (
   EXISTS(SELECT 1 FROM tls209_appln_ipc i WHERE i.appln_id=a.appln_id AND(
     i.ipc_class_symbol LIKE 'G06N%'OR i.ipc_class_symbol LIKE 'G06V%'OR i.ipc_class_symbol LIKE 'G06F  18/%'
     OR i.ipc_class_symbol LIKE 'G06F  40/%'OR i.ipc_class_symbol LIKE 'G10L  13/%'OR i.ipc_class_symbol LIKE '
     G10L  15/%'
     OR i.ipc_class_symbol LIKE 'G10L  17/%'OR i.ipc_class_symbol LIKE 'G10L  25/%'))
   OR EXISTS(SELECT 1 FROM tls224_appln_cpc c WHERE c.appln_id=a.appln_id AND(
     c.cpc_class_symbol LIKE 'G06N%'OR c.cpc_class_symbol LIKE 'G06V%'OR c.cpc_class_symbol LIKE 'G06F  18/%'
     OR c.cpc_class_symbol LIKE 'G06F  40/%'OR c.cpc_class_symbol LIKE 'G10L  13/%'OR c.cpc_class_symbol LIKE '
     G10L  15/%'
     OR c.cpc_class_symbol LIKE 'G10L  17/%'OR c.cpc_class_symbol LIKE 'G10L  25/%'OR c.cpc_class_symbol LIKE '
     Y10S706%'))
 ) THEN 1 ELSE 0 END has_ai_core_code
FROM tls201_appln a
 LEFT JOIN tls202_appln_title t ON a.appln_id=t.appln_id
 LEFT JOIN tls203_appln_abstr b0 ON a.appln_id=b0.appln_id
WHERE a.earliest_filing_year BETWEEN 1990 AND 1990 AND a.granted='Y' AND(
 EXISTS(SELECT 1 FROM tls209_appln_ipc i WHERE i.appln_id=a.appln_id AND(
   i.ipc_class_symbol LIKE 'G06N%'OR i.ipc_class_symbol LIKE 'G06V%'OR i.ipc_class_symbol LIKE 'G06F  18/%'
   OR i.ipc_class_symbol LIKE 'G06F  40/%'OR i.ipc_class_symbol LIKE 'G10L  13/%'OR i.ipc_class_symbol LIKE 'G10L
       15/%'
   OR i.ipc_class_symbol LIKE 'G10L  17/%'OR i.ipc_class_symbol LIKE 'G10L  25/%'))
 OR EXISTS(SELECT 1 FROM tls224_appln_cpc c WHERE c.appln_id=a.appln_id AND(
   c.cpc_class_symbol LIKE 'G06N%'OR c.cpc_class_symbol LIKE 'G06V%'OR c.cpc_class_symbol LIKE 'G06F  18/%'
   OR c.cpc_class_symbol LIKE 'G06F  40/%'OR c.cpc_class_symbol LIKE 'G10L  13/%'OR c.cpc_class_symbol LIKE 'G10L
       15/%'
   OR c.cpc_class_symbol LIKE 'G10L  17/%'OR c.cpc_class_symbol LIKE 'G10L  25/%'OR c.cpc_class_symbol LIKE '
     Y10S706%'))
 OR EXISTS(SELECT 1 FROM tls209_appln_ipc i WHERE i.appln_id=a.appln_id AND(
   i.ipc_class_symbol LIKE 'A61B   5/00%'OR i.ipc_class_symbol LIKE 'A61B  34/00%'OR i.ipc_class_symbol LIKE '
     A63F  13/67%'
   OR i.ipc_class_symbol LIKE 'B23K  31/00%'OR i.ipc_class_symbol LIKE 'B25J%'OR i.ipc_class_symbol LIKE 'B29C
     65/00%'
   OR i.ipc_class_symbol LIKE 'B60W  30/%'OR i.ipc_class_symbol LIKE 'B60W  40/%'OR i.ipc_class_symbol LIKE 'B60W
       50/%'OR i.ipc_class_symbol LIKE 'B60W  60/%'
   OR i.ipc_class_symbol LIKE 'B62D  15/02%'OR i.ipc_class_symbol LIKE 'B62D  57/032%'OR i.ipc_class_symbol LIKE
     'B64C%'OR i.ipc_class_symbol LIKE 'B64G   1/24%'
   OR i.ipc_class_symbol LIKE 'E21B  41/00%'OR i.ipc_class_symbol LIKE 'F02D  41/14%'OR i.ipc_class_symbol LIKE '
     F03D   7/04%'OR i.ipc_class_symbol LIKE 'F16H  61/00%'
   OR i.ipc_class_symbol LIKE 'G01C  21/%'OR i.ipc_class_symbol LIKE 'G01N  29/44%'OR i.ipc_class_symbol LIKE '
     G01N  33/00%'OR i.ipc_class_symbol LIKE 'G01R  31/2%'
   OR i.ipc_class_symbol LIKE 'G01R  31/3%'OR i.ipc_class_symbol LIKE 'G01S   7/41%'OR i.ipc_class_symbol LIKE '
     G01S  13/%'OR i.ipc_class_symbol LIKE 'G01S  17/%'
   OR i.ipc_class_symbol LIKE 'G02B  27/01%'OR i.ipc_class_symbol LIKE 'G05B  13/02%'OR i.ipc_class_symbol LIKE '
     G05D   1/%'OR i.ipc_class_symbol LIKE 'G06E%'
   OR i.ipc_class_symbol LIKE 'G06F   9/44%'OR i.ipc_class_symbol LIKE 'G06F  11/%'OR i.ipc_class_symbol LIKE '
```

```
    G06F  15/%'OR i.ipc_class_symbol LIKE 'G06F  16/%'
  OR i.ipc_class_symbol LIKE 'G06F  17/%'OR i.ipc_class_symbol LIKE 'G06F  19/%'OR i.ipc_class_symbol LIKE 'G06F
      21/%'OR i.ipc_class_symbol LIKE 'G06G   7/00%'
  OR i.ipc_class_symbol LIKE 'G06J   1/00%'OR i.ipc_class_symbol LIKE 'G06K   7/14%'OR i.ipc_class_symbol LIKE '
    G06K   9/%'OR i.ipc_class_symbol LIKE 'G06Q%'
  OR i.ipc_class_symbol LIKE 'G06T%'OR i.ipc_class_symbol LIKE 'G08B  29/18%'OR i.ipc_class_symbol LIKE 'G08G%'
    OR i.ipc_class_symbol LIKE 'G09B%'
  OR i.ipc_class_symbol LIKE 'G10L%'OR i.ipc_class_symbol LIKE 'G11B  20/10%'OR i.ipc_class_symbol LIKE 'G16H%'
    OR i.ipc_class_symbol LIKE 'H01M   8/04992%'
  OR i.ipc_class_symbol LIKE 'H02H   1/00%'OR i.ipc_class_symbol LIKE 'H02P  21/00%'OR i.ipc_class_symbol LIKE '
    H02P  23/00%'OR i.ipc_class_symbol LIKE 'H03H  17/02%'
  OR i.ipc_class_symbol LIKE 'H04L  12/%'OR i.ipc_class_symbol LIKE 'H04L  25/0%'OR i.ipc_class_symbol LIKE '
    H04L  41/%'OR i.ipc_class_symbol LIKE 'H04L  51/%'
  OR i.ipc_class_symbol LIKE 'H04N  21/466%'OR i.ipc_class_symbol LIKE 'H04R  25/00%'))
 OR EXISTS(SELECT 1 FROM tls224_appln_cpc c WHERE c.appln_id=a.appln_id AND(
  c.cpc_class_symbol LIKE 'B25J%'OR c.cpc_class_symbol LIKE 'B60W  60/%'OR c.cpc_class_symbol LIKE 'B62D
    57/032%'OR c.cpc_class_symbol LIKE 'G01C  21/%'
  OR c.cpc_class_symbol LIKE 'G02B  27/01%'OR c.cpc_class_symbol LIKE 'G05B  13/02%'OR c.cpc_class_symbol LIKE '
    G05B2219/40%'OR c.cpc_class_symbol LIKE 'G05D   1/%'
  OR c.cpc_class_symbol LIKE 'G06F  15/18%'OR c.cpc_class_symbol LIKE 'G06F  16/%'OR c.cpc_class_symbol LIKE '
    G06F  17/16%'OR c.cpc_class_symbol LIKE 'G06F  17/2%'
  OR c.cpc_class_symbol LIKE 'G06F  17/30%'OR c.cpc_class_symbol LIKE 'G06K   9/%'OR c.cpc_class_symbol LIKE '
    G06T   1/20%'OR c.cpc_class_symbol LIKE 'G06T   7/00%'
  OR c.cpc_class_symbol LIKE 'G06T2207/20%'OR c.cpc_class_symbol LIKE 'G08G   1/16%'OR c.cpc_class_symbol LIKE '
    H04L  51/02%'OR c.cpc_class_symbol LIKE 'G16H%'
  OR c.cpc_class_symbol LIKE 'Y10S901%'))
);
```

Listing 1: PATSTAT AI patent retrieval SQL used for the AI patent corpus

B Physical AI and Physical AI Risk Filtering Rules

Listing 2 documents the SQL-equivalent filtering rules used to move from the AI corpus to the Physical AI candidate set and then to the Physical AI risk subset. The production extraction script implements the same term logic; this listing records it in a database-readable form for reproducibility.

```
-- SQL-equivalent reproducibility specification for AI -> Physical AI -> Physical AI Risks
-- This file documents the filtering logic used after AI-corpus retrieval.
-- The actual local extraction was implemented in:
--   ../50_physical_ai_dataset/04_scripts/build_physical_ai_dataset.py
-- and audited in:
--   ../50_physical_ai_dataset/05_logs/physical_ai_dataset_manifest.json
--   ../50_physical_ai_dataset/03_audit/physical_ai_collection_status_20260627.md

-- ------------------------------------------------------------------------
-- Stage 1. Physical AI selection from AI corpus
-- Select if either:
--   A. a strong Physical AI marker appears in title/abstract/keywords/classification text; or
--   B. a broad perception/control term appears together with a physical/embodied/robotic context marker.
-- ------------------------------------------------------------------------

WITH normalized_ai_corpus AS (
  SELECT
    source_family,
    source_id,
    year,
    LOWER(CONCAT_WS(' ',
      title,
      abstract,
      keywords,
      l1_label,
      l2_label,
      l3_label,
      ipc,
      cpc,
      source_metadata
    )) AS doc_text
  FROM ai_corpus
```

```
),
physical_ai_candidates AS (
  SELECT *
  FROM normalized_ai_corpus
  WHERE
    (
      doc_text LIKE '%physical ai%'
      OR doc_text LIKE '%physical artificial intelligence%'
      OR doc_text LIKE '%embodied ai%'
      OR doc_text LIKE '%embodied artificial intelligence%'
      OR doc_text LIKE '%embodied agent%'
      OR doc_text LIKE '%embodied agents%'
      OR doc_text LIKE '%embodied intelligence%'
      OR doc_text LIKE '%robot%'
      OR doc_text LIKE '%robotic%'
      OR doc_text LIKE '%robotics%'
      OR doc_text LIKE '%humanoid%'
      OR doc_text LIKE '%autonomous vehicle%'
      OR doc_text LIKE '%autonomous vehicles%'
      OR doc_text LIKE '%autonomous driving%'
      OR doc_text LIKE '%autonomous car%'
      OR doc_text LIKE '%autonomous mobility%'
      OR doc_text LIKE '%autonomous system%'
      OR doc_text LIKE '%autonomous systems%'
      OR doc_text LIKE '%autonomous weapons%'
      OR doc_text LIKE '%self-driving%'
      OR doc_text LIKE '%drone%'
      OR doc_text LIKE '%uav%'
      OR doc_text LIKE '%unmanned%'
      OR doc_text LIKE '%cyber-physical%'
      OR doc_text LIKE '%cyber physical%'
      OR doc_text LIKE '%cps%'
      OR doc_text LIKE '%digital twin%'
      OR doc_text LIKE '%digital twins%'
      OR doc_text LIKE '%sensor fusion%'
      OR doc_text LIKE '%simultaneous localization and mapping%'
      OR doc_text LIKE '%simultaneous localisation and mapping%'
      OR doc_text LIKE '%slam%'
      OR doc_text LIKE '%sim-to-real%'
      OR doc_text LIKE '%sim to real%'
      OR doc_text LIKE '%robotic manipulation%'
      OR doc_text LIKE '%manipulation%'
      OR doc_text LIKE '%grasping%'
      OR doc_text LIKE '%gripper%'
      OR doc_text LIKE '%navigation%'
      OR doc_text LIKE '%motion planning%'
      OR doc_text LIKE '%path planning%'
      OR doc_text LIKE '%mobile robot%'
      OR doc_text LIKE '%service robot%'
      OR doc_text LIKE '%industrial robot%'
      OR doc_text LIKE '%collaborative robot%'
      OR doc_text LIKE '%cobot%'
      OR doc_text LIKE '%machine vision%'
      OR doc_text LIKE '%smart manufacturing%'
      OR doc_text LIKE '%smart factory%'
      OR doc_text LIKE '%factory automation%'
      OR doc_text LIKE '%industrial automation%'
      OR doc_text LIKE '%actuator%'
      OR doc_text LIKE '%actuation%'
      OR doc_text LIKE '%robot control%'
      OR doc_text LIKE '%vehicle control%'
      OR doc_text LIKE '%physical safety%'
      OR doc_text LIKE '%피지컬 ai%'
      OR doc_text LIKE '%피지컬 인공지능%'
      OR doc_text LIKE '%물리 인공지능%'
      OR doc_text LIKE '%물리적 인공지능%'
      OR doc_text LIKE '%체화 인공지능%'
      OR doc_text LIKE '%체화형 인공지능%'
      OR doc_text LIKE '%임바디드%'
      OR doc_text LIKE '%로봇%'
      OR doc_text LIKE '%로보틱스%'
      OR doc_text LIKE '%휴머노이드%'
      OR doc_text LIKE '%자율주행%'
```

```
      OR doc_text LIKE '%자율 주행%'
      OR doc_text LIKE '%자율주행차%'
      OR doc_text LIKE '%자율 차량%'
      OR doc_text LIKE '%자율무기%'
      OR doc_text LIKE '%자율 무기%'
      OR doc_text LIKE '%무인%'
      OR doc_text LIKE '%드론%'
      OR doc_text LIKE '%사이버 물리%'
      OR doc_text LIKE '%사이버-물리%'
      OR doc_text LIKE '%디지털 트윈%'
      OR doc_text LIKE '%센서 융합%'
      OR doc_text LIKE '%동시 위치%'
      OR doc_text LIKE '%지도 작성%'
      OR doc_text LIKE '%로봇 조작%'
      OR doc_text LIKE '%파지%'
      OR doc_text LIKE '%그리퍼%'
      OR doc_text LIKE '%내비게이션%'
      OR doc_text LIKE '%네비게이션%'
      OR doc_text LIKE '%동작 계획%'
      OR doc_text LIKE '%경로 계획%'
      OR doc_text LIKE '%이동 로봇%'
      OR doc_text LIKE '%서비스 로봇%'
      OR doc_text LIKE '%산업용 로봇%'
      OR doc_text LIKE '%협동로봇%'
      OR doc_text LIKE '%협동 로봇%'
      OR doc_text LIKE '%기계 비전%'
      OR doc_text LIKE '%스마트팩토리%'
      OR doc_text LIKE '%스마트 팩토리%'
      OR doc_text LIKE '%공장 자동화%'
      OR doc_text LIKE '%제조 자동화%'
      OR doc_text LIKE '%액추에이터%'
      OR doc_text LIKE '%물리적 안전%'
    )
    OR
    (
      (
        doc_text LIKE '%computer vision%'
        OR doc_text LIKE '%object detection%'
        OR doc_text LIKE '%object recognition%'
        OR doc_text LIKE '%image recognition%'
        OR doc_text LIKE '%image segmentation%'
        OR doc_text LIKE '%semantic segmentation%'
        OR doc_text LIKE '%instance segmentation%'
        OR doc_text LIKE '%scene understanding%'
        OR doc_text LIKE '%action recognition%'
        OR doc_text LIKE '%visual perception%'
        OR doc_text LIKE '%vision-language%'
        OR doc_text LIKE '%vision language%'
        OR doc_text LIKE '%visual navigation%'
        OR doc_text LIKE '%remote sensing%'
        OR doc_text LIKE '%reinforcement learning%'
        OR doc_text LIKE '%safe reinforcement learning%'
        OR doc_text LIKE '%control%'
        OR doc_text LIKE '%optimal control%'
        OR doc_text LIKE '%sensor%'
        OR doc_text LIKE '%sensing%'
        OR doc_text LIKE '%perception%'
        OR doc_text LIKE '%컴퓨터비전%'
        OR doc_text LIKE '%컴퓨터 비전%'
        OR doc_text LIKE '%객체탐지%'
        OR doc_text LIKE '%객체 탐지%'
        OR doc_text LIKE '%객체인식%'
        OR doc_text LIKE '%객체 인식%'
        OR doc_text LIKE '%이미지 인식%'
        OR doc_text LIKE '%영상인식%'
        OR doc_text LIKE '%영상 인식%'
        OR doc_text LIKE '%이미지 분할%'
        OR doc_text LIKE '%영상 분할%'
        OR doc_text LIKE '%장면 이해%'
        OR doc_text LIKE '%행동 인식%'
        OR doc_text LIKE '%강화학습%'
        OR doc_text LIKE '%강화 학습%'
        OR doc_text LIKE '%제어%'
```

```
        OR doc_text LIKE '%센서%'
        OR doc_text LIKE '%감지%'
        OR doc_text LIKE '%인지%'
      )
      AND
      (
        doc_text LIKE '%physical%'
        OR doc_text LIKE '%embodied%'
        OR doc_text LIKE '%robot%'
        OR doc_text LIKE '%robotic%'
        OR doc_text LIKE '%robotics%'
        OR doc_text LIKE '%humanoid%'
        OR doc_text LIKE '%autonomous%'
        OR doc_text LIKE '%vehicle%'
        OR doc_text LIKE '%driving%'
        OR doc_text LIKE '%drone%'
        OR doc_text LIKE '%uav%'
        OR doc_text LIKE '%unmanned%'
        OR doc_text LIKE '%mobility%'
        OR doc_text LIKE '%cyber-physical%'
        OR doc_text LIKE '%cyber physical%'
        OR doc_text LIKE '%digital twin%'
        OR doc_text LIKE '%manufacturing%'
        OR doc_text LIKE '%factory%'
        OR doc_text LIKE '%industrial%'
        OR doc_text LIKE '%machine vision%'
        OR doc_text LIKE '%sensor fusion%'
        OR doc_text LIKE '%slam%'
        OR doc_text LIKE '%navigation%'
        OR doc_text LIKE '%manipulation%'
        OR doc_text LIKE '%grasp%'
        OR doc_text LIKE '%actuator%'
        OR doc_text LIKE '%물리%'
        OR doc_text LIKE '%체화%'
        OR doc_text LIKE '%임바디%'
        OR doc_text LIKE '%로봇%'
        OR doc_text LIKE '%휴머노이드%'
        OR doc_text LIKE '%자율%'
        OR doc_text LIKE '%차량%'
        OR doc_text LIKE '%주행%'
        OR doc_text LIKE '%드론%'
        OR doc_text LIKE '%무인%'
        OR doc_text LIKE '%모빌리티%'
        OR doc_text LIKE '%사이버 물리%'
        OR doc_text LIKE '%디지털 트윈%'
        OR doc_text LIKE '%제조%'
        OR doc_text LIKE '%공장%'
        OR doc_text LIKE '%산업%'
        OR doc_text LIKE '%센서 융합%'
        OR doc_text LIKE '%내비게이션%'
        OR doc_text LIKE '%네비게이션%'
        OR doc_text LIKE '%조작%'
        OR doc_text LIKE '%파지%'
        OR doc_text LIKE '%액추에이터%'
      )
    )
)

-- ------------------------------------------------------------------------
-- Stage 2. Physical AI Risk subset from Physical AI candidates
-- Apply risk/safety/security/ethics/failure terms to title/abstract/keyword text.
-- This produces the Physical AI risk subset used for L4 risk-card generation.
-- ------------------------------------------------------------------------

SELECT *
FROM physical_ai_candidates
WHERE
  doc_text LIKE '%risk%'
  OR doc_text LIKE '%risks%'
  OR doc_text LIKE '%safety%'
  OR doc_text LIKE '%safe%'
  OR doc_text LIKE '%unsafe%'
  OR doc_text LIKE '%harm%'
```

```
  OR doc_text LIKE '%hazard%'
  OR doc_text LIKE '%accident%'
  OR doc_text LIKE '%failure%'
  OR doc_text LIKE '%fault%'
  OR doc_text LIKE '%error%'
  OR doc_text LIKE '%robustness%'
  OR doc_text LIKE '%vulnerability%'
  OR doc_text LIKE '%threat%'
  OR doc_text LIKE '%attack%'
  OR doc_text LIKE '%security%'
  OR doc_text LIKE '%cybersecurity%'
  OR doc_text LIKE '%privacy%'
  OR doc_text LIKE '%ethics%'
  OR doc_text LIKE '%ethical%'
  OR doc_text LIKE '%bias%'
  OR doc_text LIKE '%discrimination%'
  OR doc_text LIKE '%accountability%'
  OR doc_text LIKE '%liability%'
  OR doc_text LIKE '%transparency%'
  OR doc_text LIKE '%explainability%'
  OR doc_text LIKE '%trust%'
  OR doc_text LIKE '%misinformation%'
  OR doc_text LIKE '%misalignment%'
  OR doc_text LIKE '%jailbreak%'
  OR doc_text LIKE '%prompt injection%'
  OR doc_text LIKE '%안전%'
  OR doc_text LIKE '%위험%'
  OR doc_text LIKE '%리스크%'
  OR doc_text LIKE '%위해%'
  OR doc_text LIKE '%유해%'
  OR doc_text LIKE '%사고%'
  OR doc_text LIKE '%실패%'
  OR doc_text LIKE '%오류%'
  OR doc_text LIKE '%결함%'
  OR doc_text LIKE '%취약%'
  OR doc_text LIKE '%공격%'
  OR doc_text LIKE '%보안%'
  OR doc_text LIKE '%프라이버시%'
  OR doc_text LIKE '%윤리%'
  OR doc_text LIKE '%편향%'
  OR doc_text LIKE '%차별%'
  OR doc_text LIKE '%책임%'
  OR doc_text LIKE '%투명%'
  OR doc_text LIKE '%설명가능%'
  OR doc_text LIKE '%신뢰%'
  OR doc_text LIKE '%허위%'
  OR doc_text LIKE '%탈옥%'
  OR doc_text LIKE '%프롬프트 주입%';
```

Listing 2: SQL-equivalent Physical AI and Physical AI Risk filtering rules

C Web of Science Search Query

Listing 3 reproduces the Web of Science topic-search query preserved in the raw source collection. The listing below is imported verbatim so that the query text remains unchanged.

```
TS=("artificial intelligence" OR AI OR "A.I." OR "machine learning" OR "statistical learning" OR "supervised
     learning" OR "unsupervised learning" OR "semi-supervised learning" OR "self-supervised learning" OR "
     reinforcement learning" OR "transfer learning" OR "federated learning" OR "deep learning" OR "neural network
     " OR "artificial neural network" OR DNN OR CNN OR RNN OR LSTM OR GAN OR GNN OR "convolutional neural network
     " OR "recurrent neural network" OR "graph neural network" OR "natural language processing" OR NLP OR "
     language model" OR "large language model" OR LLM OR "foundation model" OR transformer OR "attention
     mechanism" OR "machine translation" OR "information extraction" OR "question answering" OR "text mining" OR
     "sentiment analysis" OR chatbot OR "conversational AI" OR "generative AI" OR GenAI OR "generative artificial
      intelligence" OR "generative model" OR "diffusion model" OR "generative adversarial network" OR "
     variational autoencoder" OR "multimodal AI" OR "multimodal model" OR "frontier AI" OR "artificial general
     intelligence" OR AGI OR "computer vision" OR "image recognition" OR "object detection" OR "image
     segmentation" OR "facial recognition" OR "pattern recognition" OR "speech recognition" OR "speech synthesis"
      OR "speech processing" OR "speaker recognition" OR "expert system" OR "knowledge-based system" OR "
```

```
knowledge representation" OR "inference engine" OR "automated reasoning" OR "rule-based system" OR "logic
programming" OR fuzzy OR "fuzzy logic" OR "Bayesian network" OR "genetic algorithm" OR "evolutionary
algorithm" OR "swarm intelligence" OR "recommender system" OR "predictive analytics" OR "decision support"
OR "anomaly detection" OR "automated classification" OR "automated decision-making" OR "intelligent decision
-making" OR "autonomous driving" OR "autonomous vehicle" OR "unmanned system" OR "intelligent robot" OR "
robot intelligence" OR "intelligent automation" OR "data mining" OR "인공지능" OR "인공 지능" OR "에이아이" OR
 "지능정보" OR "지능정보기술" OR "지능정보사회" OR "지능형 시스템" OR "지능형 서비스" OR "지능형 에이전트" OR "지
능형 정보시스템" OR "기계학습" OR "머신러닝" OR "지도학습" OR "비지도학습" OR "준지도학습" OR "자기지도학습" OR "
강화학습" OR "전이학습" OR "연합학습" OR "딥러닝" OR "심층학습" OR "신경망" OR "인공신경망" OR "합성곱 신경망" OR
 "순환 신경망" OR "그래프 신경망" OR "자연어처리" OR "자연어 처리" OR "언어모델" OR "대규모 언어모델" OR "거대언어
모델" OR "기반모델" OR "파운데이션 모델" OR "트랜스포머" OR "어텐션" OR "기계번역" OR "정보추출" OR "질의응답" OR
 "텍스트마이닝" OR "감성분석" OR "챗봇" OR "대화형 AI" OR "생성형 AI" OR "생성 AI" OR "생성형 인공지능" OR "생성
 모델" OR "확산모델" OR "적대적 생성 신경망" OR "변분 오토인코더" OR "멀티모달 AI" OR "멀티모달 모델" OR "프론티
어 AI" OR "범용 인공지능" OR "컴퓨터비전" OR "컴퓨터 비전" OR "영상인식" OR "이미지인식" OR "객체인식" OR "객체탐
지" OR "객체검출" OR "영상분할" OR "이미지분할" OR "얼굴인식" OR "패턴인식" OR "음성인식" OR "음성합성" OR "음성
처리" OR "화자인식" OR "전문가시스템" OR "전문가 시스템" OR "지식기반시스템" OR "지식기반 시스템" OR "지식표현" OR
 "추론엔진" OR "자동추론" OR "규칙기반 시스템" OR "논리프로그래밍" OR "퍼지" OR "퍼지논리" OR "베이지안 네트워크"
OR "유전 알고리즘" OR "진화 알고리즘" OR "군집지능" OR "추천시스템" OR "예측분석" OR "의사결정지원" OR "이상탐지"
 OR "자동분류" OR "자동화된 의사결정" OR "지능형 의사결정" OR "자율주행" OR "자율주행차" OR "무인시스템" OR "지능
형 로봇" OR "로봇지능" OR "지능형 자동화" OR "데이터마이닝") AND CU=(South Korea) AND PY=(1990-2026) AND DT=(
Article OR Proceedings Paper)
```

Listing 3: Web of Science topic-search query preserved from the local source collection

D Data Availability: Full Construction Corpora

This appendix documents the full-corpus datasets that underlie the integrated counts in Table 3. Master and final datasets are treated as read-only. For every source family a *domestic/global comparison key* (Korea vs. worldwide total) is recorded, so that all downstream statistics can be reported as domestic, overseas, and total.

Table 12: Full construction corpus: patents (PATSTAT)

Dataset	Records	Note
AI patents (integrated)	2,032,236	Table 3 PATSTAT AI corpus
Non-AI patents (full)	3,850,424	
AI Infrastructure patents (integrated)	522,085	Table 3 PATSTAT AI infrastructure corpus
Total patent universe (AI + Non-AI)	**5,887,803**	Analysis base rate / denominator

Note: PATSTAT Stage-2 joined records. Domestic/global key: applicant country.

Table 13: Full construction corpus: scholarly papers (Web of Science)

Category	Raw entries	After dedup	Note
AI papers	–	1,749,159	Table 3 integrated scholarly AI corpus
Physical AI Risks papers	–	50,375	Table 3 integrated scholarly risk corpus
AI Governance	12,101	12,101	societal/governance evidence
Manual SavedRecs	28,000	27,000	manually curated records

Note: Domestic/global key: Korea affiliation / country.

Within-category deduplication is applied; cross-category duplication (e.g., a paper appearing in both AI Global and AI Risk) is intentionally retained as a crosswalk signal. Raw `.bib` inputs comprise 2,716 top-level files and 2,701,711 BibTeX entries (with duplicates).

The current integrated record-level datasets are organized by source family and category, with headline denominators of 383,149 Physical AI records and 82,971 Physical AI risk records, split across papers, policy reports, and patents as shown in Table 3.

Table 14: Full construction corpus: policy and governance reports

Dataset	Records	Note
Curated policy/report master	4,947	largest curated master
Taxonomy/embedding training view	3,119	
Core policy reports	1,052	
Reference-chain integrated master	3,341	
Report-to-report reference edges	243	citation graph
Broader materials master (share)	124,881	domestic + overseas policy/materials
AI master dataframe (base)	105,283	
Non-AI / metadata-only policy	192,682	

Note: Domestic/global key: domestic/foreign flag.

Identifier and comparison keys. Records are keyed by application identifier (patents), WoS ID / DOI / title–year (papers), and a stable record identifier with title/year/institution (policy). Deduplication is applied within each document type; cross-type duplication is retained to support paper–policy–patent crosswalks. Every statistic derived from these corpora should be reported with a domestic (Korea) versus overseas (worldwide total) breakdown using the domestic/global key listed for each source family.

E Released Snapshot and File Inventory

At the repository level, the current Physical AI Risk Taxonomy release contains 182 L4 cards and 360 evidence/reference rows exported from the final HTML. The local repository size is approximately 26 MB, while the broader risk coevolution workspace and Web of Science export occupy approximately 4.5 GB and 12 GB, respectively.

Table 15: Released taxonomy snapshot

Layer or artifact	Count	Notes
L2 categories	3	P2, I2, S2
L3 sub-categories	24	5 system, 10 interaction, 9 societal
L4 risk cards	182	Final card set after semantic merging and review
Evidence/reference rows	360	Extracted from final HTML
Linked references	353	Rows with active reference links
Unlinked evidence rows	7	Internal evidence notes retained in the snapshot
Maximum references per card	5	Enforced curation cap

Note: Counts are exported from the final HTML release.

References

[1] NVIDIA. *NVIDIA Cosmos: Physical AI with World Foundation Models*. Accessed 2026-07-03. https://www.nvidia.com/en-us/ai/cosmos/.

[2] Sermanet, P., Majumdar, A., Irpan, A., Kalashnikov, D., and Sindhwani, V. *Generating Robot Constitutions & Benchmarks for Semantic Safety*. arXiv:2503.08663, 2025. https://arxiv.org/abs/2503.08663.

[3] ASIMOV Benchmark. *ASIMOV Benchmark v2*. Accessed 2026-07-03. https://asimov-benchmark.github.io/v2/.

[4] Chan, A., Salganik, R., Markelius, A., et al. *Harms from Increasingly Agentic Algorithmic Systems*. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). arXiv:2302.10329, 2023. doi:10.1145/3593013.3594033. https://arxiv.org/abs/2302.10329.

[5] Slattery, P., Saeri, A. K., Grundy, E. A. C., et al. *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence*. arXiv:2408.12622, 2024. https://arxiv.org/abs/2408.12622.

[6] Weidinger, L., Mellor, J., Rauh, M., et al. *Ethical and social risks of harm from Language Models*. arXiv:2112.04359, 2021. https://arxiv.org/abs/2112.04359.

[7] Asimov, I. *I, Robot*. Gnome Press, 1950 (the Three Laws first appear in "Runaround," *Astounding Science Fiction*, 1942).

[8] Murphy, R., and Woods, D. D. *Beyond Asimov: The Three Laws of Responsible Robotics*. IEEE Intelligent Systems, 24(4), 14–20, 2009. doi:10.1109/MIS.2009.69.

[9] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. *Deep Reinforcement Learning from Human Preferences*. Advances in Neural Information Processing Systems (NeurIPS), 2017. arXiv:1706.03741. https://arxiv.org/abs/1706.03741.

[10] Ouyang, L., Wu, J., Jiang, X., et al. *Training Language Models to Follow Instructions with Human Feedback*. Advances in Neural Information Processing Systems (NeurIPS), 2022. arXiv:2203.02155. https://arxiv.org/abs/2203.02155.

[11] Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[12] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016. https://arxiv.org/abs/1606.06565.

[13] Casper, S., Davies, X., Shi, C., et al. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. Transactions on Machine Learning Research (TMLR), 2023. arXiv:2307.15217. https://arxiv.org/abs/2307.15217.

[14] Askell, A., Bai, Y., Chen, A., et al. *A General Language Assistant as a Laboratory for Alignment*. arXiv:2112.00861, 2021. https://arxiv.org/abs/2112.00861.

[15] Bai, Y., Kadavath, S., Kundu, S., et al. *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073, 2022. https://arxiv.org/abs/2212.08073.

[16] OpenAI. *Preparedness Framework (Version 2)*. OpenAI, 2025. https://openai.com/index/updating-our-preparedness-framework/.

[17] Anthropic. *Anthropic's Responsible Scaling Policy*. Anthropic, 2023–2025. https://www.anthropic.com/responsible-scaling-policy.

[18] Google DeepMind. *Frontier Safety Framework (Version 2.0)*. Google DeepMind, 2025. https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/.

[19] International Organization for Standardization. *ISO 21448:2022 — Road vehicles: Safety of the intended functionality (SOTIF)*. ISO, 2022. https://www.iso.org/standard/77490.html.

[20] Underwriters Laboratories. *ANSI/UL 4600: Standard for Evaluation of Autonomous Products*. UL Standards, 2023. https://www.shopulstandards.com/ProductDetail.aspx?productId=UL4600.

[21] Achiam, J., Held, D., Tamar, A., and Abbeel, P. *Constrained Policy Optimization*. Proceedings of the 34th International Conference on Machine Learning (ICML), 2017. arXiv:1705.10528. https://arxiv.org/abs/1705.10528.

[22] Hendrycks, D., Mazeika, M., and Woodside, T. *An Overview of Catastrophic AI Risks*. arXiv:2306.12001, 2023. https://arxiv.org/abs/2306.12001.

[23] Altman, E. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999.

[24] Zupic, I., and Čater, T. *Bibliometric Methods in Management and Organization*. Organizational Research Methods, 18(3), 429–472, 2015. doi:10.1177/1094428114562629.

[25] van Eck, N. J., and Waltman, L. *Software survey: VOSviewer, a computer program for bibliometric mapping*. Scientometrics, 84, 523–538, 2010. doi:10.1007/s11192-009-0146-3.

[26] Aria, M., and Cuccurullo, C. *bibliometrix: An R-tool for comprehensive science mapping analysis*. Journal of Informetrics, 11(4), 959–975, 2017. doi:10.1016/j.joi.2017.08.007.

[27] Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. *Bibliometrics: The Leiden Manifesto for research metrics*. Nature, 520, 429–431, 2015. doi:10.1038/520429a.

[28] OECD. *OECD Patent Statistics Manual*. OECD Publishing, 2009. Accessed 2026-07-03. https://www.oecd.org/en/publications/2009/02/oecd-patent-statistics-manual_g1gh9fa4.html.

[29] OECD. *Intellectual Property Statistics*. Accessed 2026-07-03. https://www.oecd.org/en/data/datasets/intellectual-property-statistics.html.

[30] Tyndall, J. *AACODS Checklist for Appraising Grey Literature*. Flinders University, 2010. Accessed 2026-07-03. https://fac.flinders.edu.au/dspace/api/core/bitstreams/e94a96eb-0334-4300-8880-c836d4d9a676/content.

[31] Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., et al. *PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation*. Annals of Internal Medicine, 169(7), 467–473, 2018. doi:10.7326/M18-0850.

[32] Bommasani, R., Hudson, D. A., Adeli, E., et al. *On the Opportunities and Risks of Foundation Models*. Stanford CRFM. arXiv:2108.07258, 2021. https://arxiv.org/abs/2108.07258.

[33] Yampolskiy, R. V. *Taxonomy of Pathways to Dangerous Artificial Intelligence*. AAAI Workshop, 2016. arXiv:1511.03246. https://arxiv.org/abs/1511.03246.

[34] Critch, A., and Krueger, D. *AI Research Considerations for Human Existential Safety (ARCHES)*. arXiv:2006.04948, 2020. https://arxiv.org/abs/2006.04948.

[35] OECD. *OECD Framework for the Classification of AI Systems*. OECD Digital Economy Papers No. 323, 2022. doi:10.1787/cb6d9eca-en.

[36] OECD. *Defining AI incidents and related terms*. OECD Artificial Intelligence Papers No. 16, 2024. doi:10.1787/d1a8d965-en.

[37] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. 2023. Accessed 2026-07-03. https://www.nist.gov/itl/ai-risk-management-framework.

[38] MITRE. *ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems*. Accessed 2026-07-03. https://atlas.mitre.org/.

[39] Shevlane, T., Farquhar, S., Garfinkel, B., et al. *Model evaluation for extreme risks*. arXiv:2305.15324, 2023. https://arxiv.org/abs/2305.15324.

[40] Gabriel, I., Manzini, A., Keeling, G., et al. *The Ethics of Advanced AI Assistants*. Google DeepMind. arXiv:2404.16244, 2024. https://arxiv.org/abs/2404.16244.

[41] Shavit, Y., Agarwal, S., Brundage, M., et al. *Practices for Governing Agentic AI Systems*. OpenAI, 2023. https://openai.com/research/practices-for-governing-agentic-ai-systems.

[42] Chan, A., Ezell, C., Kaufmann, M., et al. *Visibility into AI Agents*. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). arXiv:2401.13138, 2024. doi:10.1145/3630106.3658948.

[43] Benton, J., Wagner, M., Christiansen, E., et al. *Sabotage Evaluations for Frontier Models*. Anthropic. arXiv:2410.21514, 2024. https://arxiv.org/abs/2410.21514.

[44] Clarivate. *Web of Science Platform*. Accessed 2026-07-03. https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/.

[45] OpenAlex. *OpenAlex Developers: Overview*. Accessed 2026-07-03. https://developers.openalex.org/.

[46] European Patent Office. *PATSTAT: EPO Worldwide Patent Statistical Database*. Accessed 2026-07-03. https://www.epo.org/en/service-support/faq/searching-patents/about-patent-information/where-can-i-find-patent-statistics.

[47] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. Findings of the Association for Computational Linguistics: ACL 2024. https://aclanthology.org/2024.findings-acl.137/. Model card: https://huggingface.co/BAAI/bge-m3.

[48] ISO/IEC. *ISO/IEC 23894:2023, Artificial intelligence – Guidance on risk management*. 2023. Accessed 2026-07-03. https://www.iso.org/standard/77304.html.

[49] ISO. *ISO 10218-1:2025, Robotics – Safety requirements – Part 1: Industrial robots*. 2025. Accessed 2026-07-03. https://www.iso.org/standard/73933.html.

[50] ISO/IEC. *ISO/IEC TR 5469:2024, Artificial intelligence – Functional safety and AI systems*. 2024. Accessed 2026-07-03. https://www.iso.org/standard/81283.html.

[51] ISO/IEC. *ISO/IEC TS 8200:2024, Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems*. 2024. Accessed 2026-07-03. https://www.iso.org/standard/83012.html.

[52] McInnes, L., Healy, J., and Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426, 2018. https://arxiv.org/abs/1802.03426.

[53] Campello, R. J. G. B., Moulavi, D., and Sander, J. *Density-Based Clustering Based on Hierarchical Density Estimates*. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2013. doi:10.1007/978-3-642-37456-2_14.

[54] McInnes, L., Healy, J., and Astels, S. *hdbscan: Hierarchical density based clustering*. Journal of Open Source Software, 2(11), 205, 2017. https://www.theoj.org/joss-papers/joss.00205/10.21105.joss.00205.pdf.

[55] Grootendorst, M. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794, 2022. https://arxiv.org/abs/2203.05794.

[56] Dempster, A. P., Laird, N. M., and Rubin, D. B. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1–22, 1977. doi:10.1111/j.2517-6161.1977.tb01600.x.

[57] Dhillon, I. S., and Modha, D. S. *Concept Decompositions for Large Sparse Text Data Using Clustering*. Machine Learning, 42(1–2), 143–175, 2001. doi:10.1023/A:1007612920971.

[58] Rousseeuw, P. J. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53–65, 1987. doi:10.1016/0377-0427(87)90125-7.

[59] von Luxburg, U. *Clustering Stability: An Overview*. Foundations and Trends in Machine Learning, 2(3), 235–274, 2010. doi:10.1561/2200000008.

[60] Kaplan, S., and Garrick, B. J. *On the Quantitative Definition of Risk*. Risk Analysis, 1(1), 11–27, 1981. doi:10.1111/j.1539-6924.1981.tb01350.x.

[61] Collins, S., Ruina, A., Tedrake, R., and Wisse, M. *Efficient Bipedal Robots Based on Passive-Dynamic Walkers*. Science, 307(5712), 1082–1085, 2005. doi:10.1126/science.1107799.

[62] Anderson, P., Chang, A., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., and Zamir, A. R. *On Evaluation of Embodied Navigation Agents*. arXiv:1807.06757, 2018. https://arxiv.org/abs/1807.06757.