Self-Harm and AI

An Incident Registry, Taxonomy, and Governance Framework

Brian J. Roberts

DATE
March 25, 2026

VERSION
2.0

MANUSCRIPT FOCUS

This paper presents a versioned, auditable registry of alleged AI–self-harm intersections together with a governance-oriented taxonomy for trajectory-structured harm.

Version note. Quantitative results, retained-corpus counts, and policy-crosswalk totals in this version are frozen as of March 4, 2026. Materials published after that date are outside the prespecified corpus and are not incorporated into the reported counts unless stated.

Abstract

Background

AI safety systems usually treat self-harm as an event-level problem: one dangerous reply, one detection decision, or one crisis-routing step. The documented record assembled here suggests a different pattern. In many cases, harm accumulates across repeated conversations, multiple sessions, or persistent personalization. When governance treats a dynamic process as a static event, trajectory-level failures become difficult to see and difficult to evaluate. This paper addresses that gap by presenting a versioned, auditable registry of alleged AI–self-harm intersections and a governance-oriented taxonomy for coding them.

Methods

We conducted an iterative multi-source search of legal, academic, benchmark, journalistic, regulatory, and provider-documentation sources published between January 2017 and March 4, 2026. Records were retained when public documentation established an AI–self-harm intersection at or above Grade C under the registry rubric. The retained corpus contains 32 records: 26 `human_harm_incident` rows and 6 `evaluation_or_demonstration` rows. Each incident was coded with a 20-field schema and a 13-subcode multi-label taxonomy. We also built a derived analysis view for trajectory structure and clustering, and a parallel policy corpus covering 16 public provider documents. The release preserves machine-readable coding files, a reconstructed search log, a `screening_ledger.csv`, a policy source-capture index, an 8-incident reliability summary, and forward audit scaffolds for incident and policy second-rater completion.

Results

Among the outcome-based human-harm records, the registry contains 13 deaths, 5 injuries, and 9 harm exposures. Twenty-four retained records are A1/A2, including 20 of the 27 outcome-based human-harm rows. The `evaluation_or_demonstration` subset consists of 5 unsafe-output records and 1 hearing-based harm-exposure record. Under the primary mechanical rule, `trajectory_structured_strict = yes` appears in 19 of 32 retained records overall and in 18 of 27 human-harm rows. A broader descriptive sensitivity rule adds no new positive cases; instead, it shifts 2 human-harm rows from `no` to `indeterminate`. The companion-dependency pathway summary is supported by both A1 and A2 material, whereas the belief-consolidation and escalation-collapse summaries are currently supported mainly by A2 complaint-based records. In the 16-document policy crosswalk, 2 documents treat self-harm as a native evaluative object, none treat it as a first-class threshold-bearing frontier domain, and the highest public governance tier observed is Class II structured product safety.

Interpretation

This registry is not epidemiology. Its contribution is narrower and more operational: it provides an evidence-stratified corpus, a disciplined evidentiary framework, and a taxonomy that makes pathway-level failure legible. Taken together, the incident registry, the reporting-infrastructure gap, and the policy crosswalk support a bounded governance inference. For this risk domain, persistent memory, affective optimization, persona continuity, and cross-session personalization may be more informative evaluation targets than product label alone. That inference depends heavily on A2 complaint-based records, but it is also reinforced by A1-documented architectural features and by the visible absence of trajectory-level fields in public reporting and frontier-governance documents.

Contents

1 Introduction

Self-harm is often governed as if it were a single event: one dangerous answer, one detection decision, and one crisis response. The cases assembled in this registry suggest a different pattern. In many documented deaths and injuries, the relevant AI interaction unfolded across repeated turns, multiple sessions, or both. Harm accumulated over time.

That distinction matters for governance. Some frontier safety frameworks assign thresholds and deployment consequences to named domains. Across sixteen public governance documents from *OpenAI*, *Anthropic*, *Google*, *xAI*, and *Meta*, self-harm does appear, but usually at the product-safety layer or an adjacent governance-post layer—through classifiers, refusal policies, crisis routing, or sensitive-conversation evaluations. In this corpus, it does not appear as a first-class frontier domain with its own thresholds and deployment consequences.

The same gap appears in public incident reporting. The *AI Incident Database* and the *OECD AI Incidents and Hazards Monitor* are organized around discrete events: one incident, one report, one set of harm descriptors. That architecture works for some harms. It is much less suited to harms that depend on repeated engagement, dependency formation, cross-session reinforcement, memory resurfacing, or slow failures to escalate during crisis.

This paper addresses that mismatch by presenting a versioned, auditable registry of alleged AI–self-harm intersections and a governance-oriented taxonomy designed to make trajectory-level failure signatures legible, codable, and governable. The central tension is straightforward. The same features that can make these systems feel useful or engaging—persistent memory, personalized context, and continuity across sessions—are also the features most repeatedly implicated in documented harm trajectories when users are vulnerable.

The registry's strongest and most limited inference is that feature architecture may be a more informative evaluation target than product label alone. The paper does not estimate prevalence, rank platforms by risk, or adjudicate legal causation. Instead, it keeps three analytic layers separate: the registry layer records documented intersections, evidence grades, and conservative subcodes; the analysis layer derives trajectory indicators and cluster-aware sensitivity variables; and the mechanism layer remains explanatory rather than evidentiary. A further caveat runs throughout: complaint-based records often preserve longer excerpts because plaintiffs are building path-dependent theories of harm. That dynamic may increase the apparent frequency of longitudinal coding. The manuscript addresses the problem by stratifying findings by evidence grade and by separately reporting the infrastructure-level gap, which does not depend on complaint data.

The remainder of the paper proceeds in four steps. Sections 2 and 3 define the registry objects and explain how the corpus was assembled and coded. Sections 4 through 6 present the incident results, the taxonomy distribution, and the policy crosswalk. Sections 7 and 8 then interpret what those results do—and do not—support for evaluation and governance.

TABLE 1

Study architecture

Layer	Unit	Core question	Main file
Public record	Legal filings, journalism, hearings, benchmarks, provider documents	What evidence anchors exist in public form?	Preserved screened ledger, `screening_ledger.csv`, and search execution log
Incident registry layer	32 incident records	What happened, how traceable is the evidence, and which subcodes are supported?	`incident_registry_coded.csv`
Analysis layer	Derived record types, strict/broad trajectory flags, and clusters	Which rows are pathway-structured, and how sensitive are counts to traceability and clustering?	`incident_registry_analysis_view.csv`
Policy crosswalk layer	16 provider-issued governance documents	Is self-harm a native evaluative object, and if so at what governance tier within document genre?	`policy_registry_coded.csv`

2 Definitions and Scope

Because the registry separates underlying events, codable rows, and derived analytic fields, the terminology matters.

An **incident** is the underlying qualifying documented AI-self-harm event. An **incident record** is the codable registry row used in tables and code assignment. A **pathway duplicate** is a retained second record for a distinct platform or pathway exposure affecting the same harmed individual. Duplicate retention preserves auditability, but deduplicated person counts are stated explicitly when relevant. For backward compatibility, the canonical CSV still uses `incident_id` as the stable row identifier; the analysis companion adds `incident_record_id` and retains `incident_id` as a legacy alias.

The paper uses **trajectory-structured harm** for harms that accrue across turns or sessions and depend on pathway dynamics, product affordances, or cross-session accumulation rather than a single output alone. In the analysis companion, `trajectory_structured_strict` is the purely mechanical rule keyed to adjudicated `LONG-*` codes, while `trajectory_structured_broad` preserves the broader descriptive rule with an `indeterminate` state for rows where an extended pathway is described but public evidence remains too thin for conservative longitudinal coding. Main-text prevalence reporting headlines the strict field; the broad field is retained as a sensitivity analysis. The strict operational indicator for trajectory structure in this registry is the presence of one or more `LONG-*` subcodes. That indicator is an operational summary of longitudinal coding, not an external test of the concept; its prevalence within the corpus reflects the coding scheme's sensitivity to documented cross-turn and cross-session patterns, not an

independent validation of a latent construct. `trajectory_structured_flag` is retained as a backward-compatible alias of the broad field. `RTI-CO` and `LONG-DEL` are registry-layer system-behavior labels, not diagnoses, mechanism claims, or truth-adjudication devices.

For analytic clarity, the main text uses a derived `record_type` split:

- `human_harm_incident` for rows centered on a documented harmed person or pathway duplicate.
- `evaluation_or_demonstration` for benchmark, red-team, hearing-demonstration, or test-prompt rows coded as `user_type = test`.

For the 6 evaluation rows, Appendix B also uses a derived `evaluation_subtype` variable with four values: `benchmark_study`, `red_team`, `hearing_demonstration`, and `app_evaluation`. This variable is analytic only; it does not alter the canonical 20-field incident schema.

Outcome labels are used conservatively and are defined before the result tables:

- `death`: a fatality is publicly reported.
- `injury`: non-fatal physical injury, self-harm injury, or comparable acute harm such as hospitalization is publicly reported.
- `harm exposure`: documented severe risk exposure, dependency, manipulation, or self-harm-relevant pathway exposure without a coded injury or death outcome.
- `unsafe output`: benchmark, red-team, hearing-demonstration, or app-evaluation evidence of unsafe model output without a coded person-level injury or death event.

The reliability boundary is likewise explicit. Grade A1 denotes a high-traceability primary-source record such as an inquest finding, judicial order, hearing transcript, or directly inspectable evaluation study. Grade A2 denotes a formal but unadjudicated allegation record such as a filed complaint. Grade B denotes a substantial journalistic or corroborated secondary-source record with meaningful evidentiary detail but without the traceability of A1 or A2. Grade C denotes a lower-traceability public record or excerpted report where the incident remains codable but evidentiary limits are material.

The scope is English-language public documentation from January 2017 through March 4, 2026. Included systems are consumer-facing AI systems such as chatbots, companion applications, and AI-mediated surfacing or moderation systems. The registry is documentation-limited, not onset-limited: public records may appear well after the underlying event.

Date field convention. The Date field in registry tables reports the best-available public date of the interaction window. Where only a filing date, publication date, or approximate period is known, that is stated explicitly in the incident narrative. Ranges indicate documented interaction windows; single dates indicate event dates or filing dates as labeled.

3 Methods

This section explains how the registry was assembled, how records were retained, how the coding scheme was applied, and how the policy comparison corpus was built.

3.1 Search Strategy and Screening

Searches covered five source families: legal sources; academic and benchmark sources; investigative or major journalistic sources; regulatory or legislative sources; and provider-issued safety or policy materials. The legal search used PACER, CourtListener, and official state-court portals. The academic and technical search used PubMed, Google Scholar, SSRN, arXiv, and benchmark or incident repositories. The journalistic search used ProQuest, Factiva, Google News, and outlet-specific follow-up. Provider-document searches targeted public preparedness frameworks, responsible-scaling policies, system cards, model cards, model specs, usage policies, and related governance posts current through the March 4, 2026 cutoff.

Representative incident-search strings included combinations such as `("artificial intelligence" OR chatbot OR LLM OR "AI companion") AND (suicide OR self-harm OR self-injury OR overdose)` and jurisdiction-specific legal queries pairing platform names with docket terms such as `complaint`, `order`, `petition`, `wrongful death`, or `hearing transcript`. Appendix A (`Search and Screening`) records the representative source-family queries, the screening workflow, the archive-reconstructed `search_execution_log.csv`, the machine-readable `screening_ledger.csv`, and the preserved 34-row screening ledger.

Screening was documentation-limited and iterative rather than a single export from one database. The public study materials preserve the screened ledger, a machine-readable screening ledger with dedupe clusters and source-capture metadata, and maintenance notes, but not the full universe of preliminary search hits returned before incident-level screening. The retained incident corpus is therefore reproducible at the screened-record level, not at the historical raw web-search impression level. Search, screening, and deduplication were performed as a single-author pass; the public materials do not support a duplicated historical screen because no archived second screener log is available.

PRISMA-ScR informed the reporting structure as a transparency template rather than as a claim of systematic-review compliance (Tricco et al., 2018).

For included sources, the public materials record `source_capture_date` plus an `archive_reference` or local snapshot hash in `screening_ledger.csv` for incident-side screening records and in `policy_source_index.csv` for retained policy documents.

TABLE 2

Registry assembly flow

Stage	Count	Note
Candidate rows in preserved screening ledger	34	Incident-level evidence anchors advanced for codability review
Excluded at screening	2	Grade D unverifiable social-media claims
Included in preserved screening ledger	32	Retained rows before later retained-corpus maintenance
Reclassified out of codable retained corpus	1	`2025-MLP-02` retained as contextual benchmark evidence only
New codable row added before the freeze	1	`2025-GEM-01`
Final public retained corpus	32	26 `human_harm_incident` rows and 6 `evaluation_or_demonstration` rows

3.2 Eligibility, Retention, and Analytic Stratification

The retention rule separates row inclusion from subcode assignment. Inclusion required three conditions: a documented AI-self-harm intersection, attributable public evidence, and enough information to code the record's outcome and evidence fields under the registry rubric. Taxonomy subcodes were then assigned only when the public record met the relevant threshold. Inclusion in the registry is **not** contingent on guaranteed subcode assignment. One B-grade death record (`2024-GPT-02`) is retained because the documented AI-self-harm intersection is clear enough for registry inclusion, but the public excerpts are too thin for conservative subcode coding.

The main text prespecifies the following analytic hierarchy:

- the full 32-record retained registry corpus;
- the 27-row outcome-based human-harm subset (person-level death, injury, or harm exposure) as the primary descriptive subset;
- the higher-traceability human-harm subset (`A1/A2`; `n = 20`);
- the `A1/A2/B` human-harm subset (`n = 22`) as a secondary sensitivity subset;
- a human-harm sensitivity subset excluding `C`-grade mechanism-coded rows (`n = 22`).

Counting rules remain conservative. Incident records are counted at the level of unique harmed individual x platform/pathway exposure or unique evaluation/demonstration record. Raw record counts are therefore documentation counts, not counts of statistically independent events. Under these rules, the current retained corpus contains 12 minor-involved records, but 10 distinct incidents involving 11 unique minor individuals once pathway duplicates are reconciled.

The higher-traceability, complaint-inclusive retained subset (`A1/A2`) contains 24 records (8 A1 primary-source; 16 A2 complaint-based). Within the outcome-based human-harm subset, the higher-traceability human-harm subset contains 20 records. Because A2 records are formal but unadjudicated allegation material, subset composition should be considered when interpreting prevalence patterns.

3.3 Extraction, Coding, and Reliability

Each incident record was coded with the 20-field canonical schema. The taxonomy contains 13 subcodes across six categories: GEN, DET, LONG, JB, RTI, and REC/MOD. Coding is deliberately multi-label and non-mutually exclusive because the goal is to identify loci of evaluative intervention rather than force each case into a single ontology.

`evidentiary_use_level` remains in the canonical CSV as metadata, but it is not a primary reporting variable in the analyses below. The manuscript instead foregrounds evidence grade, evidence type, and the derived analysis view that contains legacy row IDs, strict and broad trajectory fields, and clustering variables.

The public reliability materials now have two distinct roles. First, they preserve the retained 8-incident independent-rater summary audit already available in earlier archived materials. Second, they add full-corpus incident and policy audit scaffolds for forward completion. The incident scaffold covers taxonomy subcodes plus `outcome_category`, `reliability_grade`, `eligibility_retention_status`, and `trajectory_structured_broad`, with explicit `A1/A2`, `B`, and `C` strata and a planned blind-to-first-pass completion rule. The resulting claim is therefore narrower than a completed full-corpus second-rater package would support: the taxonomy is operationalized and partially reliability-tested, but historical raw second-rater exports for the incident audit are not publicly available and a completed policy second-rater pass is not yet available. Appendix C (`Reliability Appendix`) and Appendix D (`Policy Reliability Appendix`) make those boundaries explicit.

3.4 Policy Crosswalk Corpus and Coding

The policy crosswalk uses one provider-issued public document as its unit of analysis rather than one incident record. Its purpose is narrow: to code what each document explicitly operationalizes, not what an organization may do elsewhere. The corpus includes provider-issued public governance documents from five provider groups that fell into predefined document genres by the March 4, 2026 policy corpus freeze: preparedness frameworks, frontier safety frameworks, responsible-scaling policies, system cards, model cards, model specs, usage policies, safety reports, and closely related governance posts. `document_type` preserves the document's specific form, while `document_genre` groups documents into `frontier_or_scaling`, `product_safety_artifact`, and `policy_or_governance_post` for within-genre comparison before provider-level rollups. In this version's coarser genre taxonomy, public compliance-framework disclosure posts are grouped under `policy_or_governance_post` when the codable public text is the disclosure post rather than the underlying framework file. When a document contains multiple safety layers, the coding records the **highest governance tier actually evidenced for self-harm in that document** and also preserves `all_self_harm_layers_present` to show every explicit self-harm layer visible in the same document.

Providers entered the policy corpus only when they had public governance documents in the predefined genres by the cutoff and either appeared in the incident registry or functioned as major frontier-model governance comparators. For corpus selection and provider-level rollups, the five provider groups are Anthropic, Google / Google

DeepMind, Meta, OpenAI, and xAI; Google and Google DeepMind are treated as one provider group with two issuing-organization labels retained at the row level. Companion-app vendors without a comparable public governance-document set by the freeze date remain out of scope for this crosswalk even when they appear in the incident registry.

The coded 16-document corpus is:

TABLE 3

Policy crosswalk corpus

ID	Org.	Type (document_type)	Date	Short title / label
ANTH-25-01	Anthropic	Blog post	2025-06-27	Claude support / companionship post
ANTH-25-02	Anthropic	Blog post	2025-08-12	Claude safeguards post
ANTH-25-03	Anthropic	Blog post	2025-12-18	User well-being post
ANTH-25-04	Anthropic	Blog post	2025-12-19	California transparency post
ANTH-26-01	Anthropic	RSP	2026-02-24	Anthropic RSP 3.0
GOOG-24-01	Google	Usage policy	2024-12-17	GenAI prohibited use policy
GDM-25-01	Google DM	FSF	2025-11	Gemini 3 Pro FSF report
GDM-25-02	Google DM	Model card	2025-12	Gemini 3 Pro card
META-23-01	Meta	Usage policy	2023-07-18	Llama 2 AUP
META-25-01	Meta	Preparedness	2025-09-24	CWM preparedness report
OAI-25-01	OpenAI	Preparedness	2025-04-15	Preparedness v2
OAI-25-02	OpenAI	System card	2025-10-27	GPT-5 sensitive conversations addendum
OAI-25-03	OpenAI	Model spec	2025-12-18	Model Spec 2025-12-18
OAI-26-01	OpenAI	System card	2026-03-03	GPT-5.3 Instant card
XAI-25-01	xAI	Model card	2025-11-17	Grok 4.1 card
XAI-25-02	xAI	FSF	2025-12-30	xAI FAIF

Appendix E (`Policy Crosswalk Appendix`) exposes the full per-document ledger, URLs, excerpts, `document_genre`, coder notes, and layered-treatment field. Appendix D (`Policy Reliability Appendix`) provides the forward second-rater scaffold. The policy crosswalk should still be described as a single-coder descriptive pass at the policy layer.

4 Incident Registry Results

This section presents the findings in the same order as the study architecture. It begins with corpus composition, then turns to the human-harm records, the evaluation and demonstration records, the derived trajectory indicators, the cluster-aware sensitivity view, and finally the cross-case themes that recur across rows.

The full retained corpus contains 32 incident records. Stratified by `record_type`, it comprises 26 human-harm incident records and 6 evaluation or demonstration records. On the outcome axis used in the analytic subsets below, 27 rows fall into the human-harm

subset because `2025-MLP-01` is a hearing-demonstration row with a `HARM_EXPOSURE` outcome. Presence in the registry indicates a documented AI-self-harm intersection, not adjudicated legal causation or a population-rate claim.

Across the retained corpus, 24 records are A1/A2 and 8 are B/C. Across the 27-row outcome-based human-harm subset, 15 involve OpenAI/ChatGPT, 5 involve Character.AI, and 7 involve other platforms. Eight provider-litigation clusters account for the majority of records. Trajectory-strict coding (`LONG-*` present) appears across multiple provider groups and is not confined to a single litigation wave, although the observed frequency is shaped by which platforms' interactions are preserved in public filings.

4.1 Human-Harm Incident Records

TABLE 4

Human-harm incident records

ID (incident_id)	Platform	Date	User	Outcome	Grade	Codes
2017-IG-01	Instagram	2017-11	Minor (14)	Death	A1	REC/MOD-AMP
2017-PIN-01	Pinterest	2017-11	Minor (14)	Death	A1	REC/MOD-AMP
2023-CAI-01	Character.AI	2023-11-08	Minor (13)	Death	A2	LONG-DEP
2023-CHA-01	Chai Research	2023-03	Adult (30)	Death	B	GEN-E, GEN-M, LONG-DEP
2024-CAI-01	Character.AI	unknown	Multiple	Injury	A2	GEN-C, LONG-DEP, REC/MOD-AMP
2024-CAI-02	Character.AI	2023-04-14-2024-02-28	Minor (14)	Death	A1	GEN-E, LONG-DEP, JB-RP
2024-CAI-03	Character.AI	2024-12-09	Minor (14)	Injury	A2	LONG-DEP
2024-GPT-01	ChatGPT	2024	Adult (23)	Exposure	C	GEN-E, LONG-DEG
2024-GPT-02	ChatGPT (GPT-4o; "Harry")	2024-11-2025	Adult (29)	Death	B	none
2025-ACC-01	Companion apps	2025-08-11	Minor (13)	Exposure	C	GEN-E
2025-CAI-01	Character.AI	2025-08-19	Minor (13)	Exposure	A2	REC/MOD-AMP, LONG-DEP, JB-RP
2025-GEM-01	Google Gemini 2.5 Pro	2025-09-29-2025-10-02	Adult (36)	Death	A2	LONG-DEL, LONG-DEP, LONG-MEM, RTI-CO, GEN-C, GEN-M, DET-FN
2025-GPT-01	ChatGPT (GPT-4o)	2025-04-11	Minor (16)	Death	A2	GEN-C, GEN-E, LONG-DEG, DET-FN
2025-GPT-02	ChatGPT (GPT-4o)	2025	Adult	Exposure	C	GEN-E, LONG-DEG
2025-GPT-03	ChatGPT (GPT-4o)	2025	Adult	Exposure	C	GEN-E, GEN-C

Table 4 continued

ID (incident_id)	Platform	Date	User	Outcome	Grade	Codes
2025-GPT-04	ChatGPT (GPT-4o)	2025-08	Adult (48)	Death	A2	RTI-CO, LONG-DEL
2025-GPT-05	ChatGPT	unknown	Adult (26)	Injury	C	RTI-CO
2025-GPT-06	ChatGPT Plus (GPT-4o)	2025-08-03	Adult (56)	Death	A2	LONG-DEP, LONG-DEL, RTI-CO
2025-GPT-07	ChatGPT Plus (GPT-4o)	2025-04-2025	Adult (48)	Exposure	A2	RTI-CO, LONG-DEL, LONG-DEP
2025-GPT-08	ChatGPT (GPT-4o)	2025-08-04	Adult (26)	Exposure	A2	GEN-C, DET-FN
2025-GPT-09	ChatGPT (GPT-4o)	2025-04-2025-07	Adult (30)	Injury	A2	RTI-CO, LONG-DEL, LONG-MEM
2025-GPT-10	ChatGPT (GPT-4o)	2025-06-01-2025-06-02	Minor (17)	Death	A2	GEN-C, GEN-M, JB-MT
2025-GPT-11	ChatGPT (GPT-4o)	2025-06-2025-08-29	Adult (32)	Injury	A2	RTI-CO, LONG-DEL, LONG-DEP, GEN-E, DET-FN
2025-GPT-12	ChatGPT	2025-07-24	Adult (23)	Death	A2	GEN-E, LONG-DEP
2025-GPT-13	ChatGPT (GPT-4o)	2025-10-08-2025-11-02	Adult (40)	Death	A2	GEN-E, GEN-C, LONG-DEP, LONG-MEM
2025-REP-01	Replika	2025	Multiple	Exposure	A2	LONG-DEP, DET-FN

4.2 Human-Harm Evidence Companion Table

TABLE 5

Human-harm evidence companion

ID	Evidence anchor (evidence_type)	Grade	Provider group (provider_cluster)	Short anchor
2017-IG-01	Coroner or inquest (coroner_inquest)	A1	Meta Instagram (meta_instagram)	Primary source file
2017-PIN-01	Coroner or inquest (coroner_inquest)	A1	Pinterest (pinterest)	Primary source file
2023-CAI-01	Court filing (court_filing)	A2	Character.AI (character_ai)	1:25-cv-02907
2023-CHA-01	Investigative journalism (investigative_journalism)	B	Chai (chai)	Investigative reporting

Table 5 continued

ID	Evidence anchor (evidence_type)	Grade	Provider group (provider_cluster)	Short anchor
2024-CAI-01	Court filing (court_filing)	A2	Character.AI (character_ai)	2:24-cv-01014
2024-CAI-02	Court filing (court_filing)	A1	Character.AI (character_ai)	6:24-cv-01903
2024-CAI-03	Court filing (court_filing)	A2	Character.AI (character_ai)	1:25-cv-01295
2024-GPT-01	Court filing (court_filing)	C	OpenAI ChatGPT (openai_chatgpt)	Primary source file
2024-GPT-02	Investigative journalism (investigative_journalism)	B	OpenAI ChatGPT (openai_chatgpt)	New York Times account
2025-ACC-01	Investigative journalism (investigative_journalism)	C	Unspecified companion apps (unspecified_companion_apps)	ABC / triple j Hack interview
2025-CAI-01	Court filing (court_filing)	A2	Character.AI (character_ai)	1:25-cv-02906
2025-GEM-01	Court filing (court_filing)	A2	Google Gemini (google_gemini)	5:26-cv-01849-VKD
2025-GPT-01	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	CGC-25-628528
2025-GPT-02	Party statement (party_statement)	C	OpenAI ChatGPT (openai_chatgpt)	SMVLC press release
2025-GPT-03	Party statement (party_statement)	C	OpenAI ChatGPT (openai_chatgpt)	SMVLC press release
2025-GPT-04	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	25STCV32379
2025-GPT-05	Investigative journalism (investigative_journalism)	C	OpenAI ChatGPT (openai_chatgpt)	ABC / triple j Hack interview
2025-GPT-06	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	CGC-25-631477
2025-GPT-07	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	25STCV32386
2025-GPT-08	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	CGC-25-630809
2025-GPT-09	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	CGC-25-630811
2025-GPT-10	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	CGC-25-630808
2025-GPT-11	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	25STCV32383
2025-GPT-12	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	25STCV32382
2025-GPT-13	Court filing (court_filing)	A2	OpenAI ChatGPT (openai_chatgpt)	Primary source file

Table 5 continued

ID	Evidence anchor (evidence_type)	Grade	Provider group (provider_cluster)	Short anchor
2025-REP-01	Regulatory filing (regulatory_filing)	A2	Replika (replika)	Primary source file

4.3 Evaluation and Demonstration Records

TABLE 6

Evaluation and demonstration records

ID (incident_id)	Platform	Subtype (evaluation_subtype)	Date	Outcome	Grade	Codes
2025-MAI-01	Meta AI	App evaluation (app_evaluation)	unknown	Unsafe output	A1	GEN-E, GEN-M, LONG-MEM
2025-MHB-01	Mental health chatbots (29 agents)	Benchmark study (benchmark_study)	unknown	Unsafe output	A1	DET-FN
2025-MLP-01	Senate Judiciary hearing	Hearing demonstration (hearing_demonstration)	2025-09-16	Exposure	A1	GEN-E, GEN-M, JB-RP
2025-MLP-03	Multi-LLM red-team (6 models)	Red-team exercise (red_team)	unknown	Unsafe output	A1	JB-AC, GEN-M, DET-FN
2025-NOM-01	Nomi AI (Glimpse AI)	App evaluation (app_evaluation)	2025-01-2025-04	Unsafe output	B	GEN-C, GEN-M, GEN-E
2025-THR-01	Therapy chatbots (multi-app evaluation)	Benchmark study (benchmark_study)	unknown	Unsafe output	A1	DET-FN, GEN-M, RTI-CO

Appendix B provides the rationale for each `evaluation_subtype` mapping. The aim is descriptive clarity, not a schema change.

4.4 Analytic-Set Summary

TABLE 7

Analytic-set summary

Analytic set	n	Death	Injury	Exposure	Unsafe output
Full retained corpus	32	13	5	9	5
Outcome-based human-harm subset	27	13	5	9	0
Evaluation or demonstration records	6	0	0	1	5
Human-harm higher-traceability subset (`A1/A2`)	20	11	4	5	0
Human-harm A1/A2/B secondary sensitivity subset	22	13	4	5	0

Twelve minor-involved records correspond to 10 distinct incidents involving 11 unique minor individuals once pathway duplicates are reconciled. Twenty of the 27 human-harm rows are A1 or A2. The `HARM_EXPOSURE` category bundles dependency allegations, manipulation exposure, and severe risk exposure without confirmed physical injury. Where finer discrimination is needed, the incident narrative and assigned subcodes provide the relevant detail. All raw record counts in the table above are documentation counts rather than counts of independent events. In this release, the `A1/A2/B` human-harm subset and the human-harm subset excluding `C`-grade mechanism-coded rows coincide numerically (`n = 22`) because all five `C`-grade human-harm rows carry mechanism coding.

4.5 Trajectory-Structured Harm Prevalence

4.5.1 Primary result: mechanical strict rule

TABLE 8

Trajectory-structured harm prevalence under the strict rule

This is the manuscript's headline trajectory result because it is mechanically derivable from adjudicated registry-layer `LONG-*` coding.

Analytic set	n	Yes	No	Indet.
Full retained corpus	32	19	13	0
Human-harm subset	27	18	9	0
Human-harm higher-traceability subset (`A1/A2`)	20	15	5	0
Human-harm A1/A2/B secondary sensitivity subset	22	16	6	0
Human-harm excluding C-coded mechanism rows	22	16	6	0

4.5.2 Sensitivity analysis: broad descriptive rule

TABLE 9

Trajectory-structured harm prevalence under the broad rule

The broad rule is retained as a descriptive sensitivity check. In this release it does not create any additional `yes` rows relative to the strict rule; it only moves two human-harm records from `no` to `indeterminate`.

Analytic set	n	Yes	No	Indet.
Full retained corpus	32	19	11	2
Human-harm subset	27	18	7	2
Human-harm higher-traceability subset (`A1/A2`)	20	15	5	0
Human-harm A1/A2/B secondary sensitivity subset	22	16	5	1
Human-harm excluding C-coded mechanism rows	22	16	5	1

4.6 Cluster-Awareness Summary

TABLE 10

Cluster-awareness summary

Measure	Raw records	Unique clusters
Person clusters	27	26
Provider clusters	27	9
Case clusters	27	25
Pathway-duplicate groups	27	26

4.7 Illustrative Cross-Case Themes

The three summaries below are narrative pathway summaries anchored in counted subcodes and the derived trajectory fields; they are **not** additional coded variables or phenotype classes.

The first recurrent pathway is **companion-dependency**, organized by `LONG-DEP` and often paired with encouragement, role-play, or recommender exposure. The clearest A1 anchor is `2024-CAI-02`, while A2 cases extend the same structure through re-contact, always-available framing, and displacement of offline support.

The second recurrent pathway is **belief-consolidation**, organized by `RTI-CO` plus `LONG-DEL`, and in some cases `LONG-MEM`. Here the central failure is not only one bad reply, but repeated reinforcement of a fixed-belief frame across time. In the current retained corpus, the richest public traces of this pattern are mostly A2.

The third recurrent pathway is **escalation-collapse**, where visible crisis cues do not trigger meaningful interruption, grounding, or mode change. `DET-FN` matters most when layered onto an existing trajectory rather than read as an isolated classifier failure.

5 Failure-Mode Taxonomy and Coding Reliability

This section shifts from case counts to the coding scheme itself. The taxonomy is heterogeneous by design because the documented cases do not fail in only one way. It classifies governance-relevant loci of failure rather than forcing the corpus into a single ontology. That design is necessary because the retained rows include output failures, detection failures, longitudinal failures, framing-based bypasses, fixed-belief non-amplification failures, and recommender or moderation pathways.

TABLE 11

Taxonomy category prevalence and governance loci

Category	Subcodes	Full corpus (n/32)	Human harm (n/27)	A1/A2 subset (n/20)	Governance locus
GEN	GEN-E, GEN-C, GEN-M	19	15	10	Output safety and self-harm response boundaries
DET	DET-FN	8	5	5	Detection-to-escalation coupling
LONG	LONG-DEP, LONG-DEG, LONG-DEL, LONG-MEM	19	18	15	Multi-session design, memory behavior, dependency, and longitudinal guardrail drift

Category	Subcodes	Full corpus (n/32)	Human harm (n/27)	A1/A2 subset (n/20)	Governance locus
JB	JB-RP, JB-MT, JB-AC	5	4	4	Framing-based bypass and adversarial evaluation
RTI	RTI-CO	8	7	6	Non-amplification in fixed-belief exchanges
REC/MOD	REC/MOD-AMP	4	4	4	Surfacing, ranking, and amplification pathways

In the outcome-based human-harm subset, `LONG-DEP` appears in 13 of 27 records, `GEN-E` in 11, `GEN-C` and `RTI-CO` in 7 each, and `LONG-DEL` in 6. In the higher-traceability human-harm subset (`A1/A2`), `LONG-DEP` appears in 12 of 20, `GEN-E`, `GEN-C`, `RTI-CO`, and `LONG-DEL` in 6 each, and `DET-FN` in 5. One retained death record remains intentionally uncoded at the subcode level because the public record is insufficient for conservative assignment.

The core ambiguity rules remain conservative. `RTI-CO` is assigned only when a discrete fixed-belief non-amplification failure is documented. `LONG-DEL` requires repeated or temporally extended reinforcement rather than a single quotable exchange. `LONG-DEP` requires more than high engagement; the record must support over-attachment, exclusivity, isolation, or erosion of offline protective factors. When evidence is insufficient, the rule is no code rather than forced coding.

TABLE 12

Reliability assets and current public status

Reliability asset	Current public status
Retained incident summary audit	Historical 8-incident agreement summary preserved from an earlier coding state; not in row-level parity with the current 2.0 released taxonomy coding.
Full-corpus incident audit matrix	Archived as a forward scaffold with `rater_1` populated and `rater_2` / `adjudicated_value` reserved for future completion
Policy second-rater surface	Archived as a forward scaffold; not yet completed

Appendix C (`Reliability Appendix`) reports the retained 8-incident summary and the current scaffold boundaries. The most interpretation-sensitive seams remain GEN-E versus no-code, LONG-DEL versus RTI-CO, and LONG-DEP versus high engagement without dependency features.

Reliability demands by evidence layer. This paper's empirical claims rest on three evidentiary layers with different reliability burdens. First, the incident taxonomy has partial inter-rater support from the retained 8-incident audit (25% of the 32-row corpus; pooled kappa = 0.80), but that audit is concentrated in A1 / A2 materials and excludes injury rows; full-corpus independent review remains incomplete. Second, the 16-document governance crosswalk codes document-level public-text properties. Because this layer remains a single-coder descriptive pass, findings such as `first_class_equivalence = yes = 0` should be read as current corpus results rather than adjudicated zeroes. Third, the infrastructure-layer claim—that the inspected AI Incident Database and OECD schemas lack explicit fields for cross-episode dependency, degradation over time, and cross-session accumulation—is directly observable from public schema definitions and does not depend on inter-rater coding.

6 Policy Crosswalk

The policy crosswalk is a second empirical dataset whose unit of analysis is one public-facing governance document rather than one incident record. Its task is narrow but important: to test whether public provider documents represent trajectory-structured self-harm as a native evaluative object and whether they elevate it into a first-class, threshold-bearing governance domain.

6.1 Main-Text Coding Definitions

TABLE 13

Main-text policy coding definitions

Term	Main-text definition
Class I	Self-harm is inside the frontier or preparedness stack, with thresholded evaluation and deployment relevance.
Class II structured product safety	Self-harm has dedicated evaluations or mitigations, but remains outside frontier thresholding.
Class III	Self-harm is handled mainly through refusals, classifiers, prohibited-content rules, or basic crisis language.
Class IV reactive incident domain	Self-harm appears mainly through post-incident patches, case response, or litigation-driven remediation.
Recorded governance tier (`risk_domain_status`)	The highest governance tier actually evidenced for self-harm in the document, or a non-tier status when self-harm is absent or cannot be defensibly classed from the public text.
Document genre (`document_genre`)	The broader comparison stratum used before any provider-level rollup.
Native evaluative object	The document names self-harm as an evaluative object and pairs it with a usable screening or evaluation procedure.
First-class equivalence	Self-harm is treated on governance terms comparable to named frontier domains such as CBRN, cyber, catastrophic misuse, or loss of control.
Threshold mechanism	The document gives self-harm a trigger, level, or escalation boundary that changes governance obligations.

Term	Main-text definition
Absent / not represented	Self-harm is not present in codable form in the document.
Ambiguous / insufficiently classifiable	The document names self-harm or an adjacent domain, but the public text does not support a defensible Class I-Class IV placement.
All self-harm layers present (`all_self_harm_layers_present`)	Every explicit governance layer in the document that carries self-harm treatment, regardless of whether it is the highest tier.
Coder confidence (`coder_confidence`)	Confidence based on document clarity, not on agreement with the organization's framing.

These definitions follow the published policy-registry manual. A document can be native and still remain Class II if self-harm receives dedicated product-safety handling but not frontier-comparable governance. In the retained 16-document corpus, all six absence-coded rows fall into `Absent / not represented`; no retained document required `Ambiguous / insufficiently classifiable`.

6.2 Incident-Reporting Infrastructure

Before examining provider-issued governance documents, it is useful to note that the gap extends to the incident-reporting layer itself. The AI Incident Database contains records of AI incidents and issues; its CSET taxonomy characterizes harms, entities, and technologies, and its GMF taxonomy analyzes failure causes (Artificial Intelligence Incident Database, n.d.-a, n.d.-b, n.d.-c). The OECD common reporting framework uses 29 criteria across eight dimensions, covering incident metadata, harm details, and whether the AI system was a direct cause, contributing factor, or otherwise involved (OECD, 2025; OECD.AI, n.d.-a, n.d.-b). Both are organized around incident records, harms, contributing factors, and metadata; neither provides explicit fields for cross-episode dependency, degradation over time, cumulative cross-session effects, or reinforcement through repeated interaction. This absence is directly observable in publicly inspectable schema definitions: it is a structural property of the reporting architecture, not an interpretive judgment requiring reliability testing. Provider-issued governance and product-safety documents are more heterogeneous. Some contain trajectory-adjacent language—references to emotional reliance, multi-turn conversations, or harmful manipulation over the course of interactions—but that vocabulary is uneven and provider-specific. These documents do not present standardized fields for cross-episode dependency, degradation over time, cumulative cross-session effects, or reinforcement through repeated interaction.

6.3 Crosswalk Findings

TABLE 14

Crosswalk findings

Public governance-document finding (N=16)	Value
Documents coded as native evaluative objects	2 documents
Documents coded as proxy-represented	2 documents
Documents coded as operationally underspecified	6 documents
Documents with no representational coverage	6 documents
Documents with first-class equivalence (`first_class_equivalence=yes`)	0 documents
Documents with no threshold mechanism (`threshold_mechanism=none`)	14 documents
Documents with no deployment consequence (`deployment_consequence=none`)	13 documents
Documents with no multi-turn evaluation (`multi_turn_evaluation=none`)	11 documents
Documents with no cross-session accumulation (`cross_session_accumulation=none`)	14 documents
Documents with no dependency attention (`dependency_attention=absent`)	12 documents
Documents with no memory or personalization attention (`memory_personalization_attention=none`)	15 documents
Documents whose recorded layer set is product safety (`all_self_harm_layers_present=product_safety`)	8 documents
Documents whose recorded layer set is trust and safety (`all_self_harm_layers_present=trust_and_safety`)	2 documents
Documents with no recorded self-harm layer (`all_self_harm_layers_present=absent`)	6 documents
Highest public governance tier observed in this corpus	Class II structured product safety

TABLE 15

Crosswalk findings by genre group

Genre group	Docs	Tier profile	Native	Any threshold	Any multi-turn
Frontier or scaling (`frontier_or_scaling`)	5	Absent / not represented in all 5 documents	0	0	0
Product-safety artifact (`product_safety_artifact`)	5	Class II in 3 documents; Class III in 2	1	2	2

Genre group	Docs	Tier profile	Native	Any threshold	Any multi-turn
Policy or governance post (`policy_or_governance_post`)	6	Class II in 2 documents; Class III in 3; Absent / not represented in 1	1	0	3

Two documents are native at the document level, but none elevate self-harm into a frontier-comparable governance domain with thresholds and deployment consequences. The comparison is clearest when stratified within `document_genre`: frontier/scaling documents are absent rather than merely ambiguous in the retained corpus, product-safety artifacts carry the strongest explicit treatment, and policy/governance posts provide the most visible multi-turn discussion without thresholding. Across the current public corpus, the highest public governance tier observed remains structured product safety rather than first-class preparedness governance; no retained document instantiates Class I or Class IV. At the provider level, every provider group in the corpus has at least one document that names self-harm explicitly. No provider group has a document coding `first_class_equivalence = yes`. The representational absence of trajectory-structured self-harm from preparedness and frontier-safety documents is consistent across all five provider groups, not concentrated in one organization. Appendix E exposes per-document dates, URLs, excerpts, coded values, layered treatment, and coder confidence so the crosswalk can be directly audited at the row level.

7 Discussion

The discussion returns to the governance question that motivated the registry. The central point is not that every documented case shares one mechanism. It is that a substantial portion of the public record is structured as a pathway rather than a single event, while the public governance and reporting tools that surround the domain remain mostly event-centered.

7.1 Why Event-Level Metrics Are Not Enough

Detection-only and refusal-only metrics remain necessary, but they are not sufficient for the harms most distinctive in this corpus. Benchmark studies already show that unsafe or inadequate responses can persist in controlled evaluations (Moore et al., 2025; Pichowicz, Kotas, & Piotrowski, 2025). The incident registry extends that concern to pathway-level failures such as dependency formation, cross-session reinforcement, memory resurfacing, and visible non-escalation during crisis. The registry therefore supports a shift in evaluative focus from isolated outputs to trajectories, especially when persistent memory, persona continuity, or repeated engagement are product features.

This does not mean that every trajectory is longitudinal in the same way, or that every case implies the same mechanism. It means the public record already contains enough pathway-structured failures to justify evaluating more than one safety object at a time: event-level outputs, product features, and multi-session pathways. That distinction is also visible in public provider materials, in which frontier preparedness and sensitive-conversation handling appear as separate governance layers (OpenAI, 2025-04-15; OpenAI, 2025-10-27; Anthropic, 2025-12-18).

7.2 Bounded Evaluation Priorities

Taken together, the registry results point to four bounded evaluation priorities.

- Evaluate multi-session trajectories, not only single turns, whenever memory or persona continuity is enabled.
- Test detection-to-action coupling: whether crisis recognition actually changes system behavior, interrupts the exchange, or routes to grounded support.
- Probe non-amplification in fixed-belief contexts, with special attention to corroboration, certainty inflation, and repeated narrative reinforcement.
- Prioritize companion-like affordances such as persistent memory, proactive re-contact, and dependency-forming interaction patterns regardless of product label alone.

These are bounded design hypotheses rather than adjudications of legal duty or comparative platform risk. The legal point remains narrow: the May 21, 2025 order in *Garcia v. Character Technologies* suggests that product-architecture theories are at least cognizable at the pleading stage, but the registry itself mainly motivates evaluation priorities rather than broader jurisprudential conclusions. More generally, base-rate limits in suicide-risk prediction research remain a reason to avoid overclaiming from detection metrics alone (Spittal et al., 2025).

7.3 Limitations

The findings should be read with several constraints in mind.

- The corpus is documentation-limited and should not be read as a population estimate.
- Many A2 records depend on partial public excerpts rather than complete platform-side logs.
- Public visibility is shaped by litigation, media attention, and platform transparency, which creates ascertainment bias.
- English-language sourcing likely underrepresents non-English incidents.
- Search, screening, and deduplication were performed as a single-reviewer pass.
- The historical raw-search universe was not preserved; the study is auditable from the screened-ledger layer forward, not from the original raw-impression layer.
- The retained incident reliability materials summarize an 8-record audit, but raw historical second-rater exports are not publicly available.
- The policy crosswalk remains a single-coder descriptive pass, although the public materials now provide a forward-completable second-rater scaffold.

- This version is frozen to the March 4, 2026 search cutoff, incident corpus freeze, and policy corpus freeze; publication and integrity-review work continued afterward.

8 Conclusion

The registry's strongest claim is disciplined and limited. It is not epidemiology, it does not estimate prevalence, and it does not rank platforms by risk. What the public record does support is a different analytic lens: some documented failures are best understood as pathway-structured harms, and event-centered evaluation is not designed to capture them well.

That finding is enough to justify memory-aware, multi-session, dependency-sensitive, and non-amplification-focused evaluation. The manuscript's contribution is therefore not a maximal policy inventory. It is a bounded empirical argument that trajectory structure is visible in the documented record, operationalized through explicit derived rules, and relevant to how AI systems should be evaluated and governed.

Appendices

The six appendices below form part of the manuscript. Standalone supplement files are also available from the downloads page.

A Search and Screening

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Release audit date: 2026-03-25
Public release date: 2026-03-25

A.1 Purpose

This supplement makes the registry build auditable at the screened-record level and documents the strongest search-history reconstruction still supportable from the archived materials. It records the source families searched, representative search strings, screening and deduplication rules, the retained screened ledger, the retained-corpus maintenance note required to reconcile the preserved 34-row screening ledger with the 32-row public registry corpus, the archive-reconstructed execution log in `search_execution_log.csv`, and the row-level `screening_ledger.csv` export.

The registry is reproducible from the screened-record layer forward. It does **not** preserve every preliminary search-engine impression, syndicated duplicate, or transient web result returned before incident-level screening. The new execution log is therefore an archive reconstruction of the search workflow, not a recovered raw-identification export from the original build materials.

A.2 Date Semantics

- `search cutoff`: March 4, 2026. No incident or policy document first identified after this date enters the retained incident corpus or retained policy corpus.
- `incident corpus freeze`: March 4, 2026. Counts in the manuscript and `incident_registry_coded.csv` are keyed to this boundary.
- `policy corpus freeze`: March 4, 2026. Counts in the manuscript and `policy_registry_coded.csv` are keyed to this boundary.
- `release audit date`: March 25, 2026. Integrity-review date for the posted study materials.
- `public release date`: March 25, 2026. Publication date for the manuscript and standalone supplement files.

A.3 Search-Execution Log Status

`search_execution_log.csv` records one row per archive-reconstructed database or document-source workflow. The file includes:

- `source_family`
- `platform_or_database`
- `run_date`
- `search_string`
- `filters`
- `sort_order`
- `results_returned_n`
- `advanced_to_screening_n`
- `notes`

`results_returned_n = not_preserved` indicates that the raw hit count was not archived in the original build materials and should not be reverse-engineered after the fact.

A.4 Screening Ledger and Source-Capture Files

Two additional CSV files now expose the screened layer more directly:

- `screening_ledger.csv` — one row per candidate advanced to screening, with `dedupe_cluster_id`, `final_disposition`, `exclusion_reason`, `source_capture_date`, `archive_reference`, and `local_snapshot_sha256` where a local source copy exists.
- `policy_source_index.csv` — one row per retained policy document with `document_genre`, `source_capture_date`, and the public archive reference used in this release.

`source_capture_date` in these files records the capture or verification date represented by the current public materials. It should not be back-interpreted as the original historical search date unless the file explicitly says so.

A.5 Source Families and Representative Searches

Source family	Coverage	Representative queries or retrieval logic	Execution window	Notes
Legal and adjudicative sources	PACER, CourtListener, official state-court portals, docket follow-up	`("artificial intelligence" OR chatbot OR LLM OR Gemini OR ChatGPT OR Character.AI) AND (suicide OR self-harm OR wrongful death OR complaint OR order)`	Iterative collection through 2026-03-04	Used for complaints, orders, petitions, and other formal filings.
Academic and benchmark sources	PubMed, Google Scholar, SSRN, arXiv, benchmark repositories	`("chatbot" OR "large language model" OR companion) AND (suicide OR self-harm OR mental health OR safety evaluation)`	Iterative collection through 2026-03-04	Advanced only when the source contained a codable benchmark, evaluation, or incident anchor.
Journalistic and investigative sources	ProQuest, Factiva, Google News, outlet follow-up	`(AI OR chatbot OR companion) AND (suicide OR self-harm OR death OR overdose OR delusion)`	Iterative collection through 2026-03-04	Used for incident discovery and corroboration; syndicated duplicates were collapsed.
Legislative and regulatory sources	Hearing archives, agency releases, legislative records	`(AI chatbot self-harm hearing)`; provider names plus `hearing`, `FTC`, `attorney general`, `petition`, `transcript`	Iterative collection through 2026-03-04	Used for hearing transcripts, agency complaints, and formal public record.
Provider-issued safety and governance materials	Preparedness frameworks, responsible-scaling policies, system cards, model cards, model specs, usage policies, governance posts	Provider-specific retrieval from official public documentation pages; one document row per codable document	Current through 2026-03-04	Used for the 16-document policy corpus reported in the manuscript.

A.6 Screening Workflow

A.6.1 Reviewer roles

- Search, screening, and deduplication were performed by the author as a single-reviewer pass.
- The public materials do not support a duplicated historical screen because no archived second screener log is available.
- The preserved materials are sufficient to audit decisions from the screened-ledger layer forward.

A.6.2 Incident inclusion rule

Advance a candidate to the codable ledger only when all three conditions hold:

1. The record documents a specific AI-self-harm intersection rather than general commentary.
2. The record is attributable to a named platform, system, or evaluation object.
3. The public evidence is sufficient to code outcome and evidence fields under the registry rubric.

A.6.3 Exclusion rule

Exclude records that are:

- unverifiable social-media claims or screenshots without attributable provenance;
- duplicates or syndicated copies without new codable evidence;
- general commentary or policy discussion without a specific incident or evaluation anchor;
- too thin to support even Grade C coding.

A.6.4 Deduplication rule

Deduplicate first at the report level, then at the incident level. Preserve distinct platform or pathway exposures when the same harmed person encountered multiple systems in separately codable ways. This is why the Molly Russell inquest yields both `2017-IG-01` and `2017-PIN-01`.

A.7 Registry Assembly Flow

Stage	Count	Note
Candidate rows in preserved screening ledger	34	Incident-level evidence anchors advanced for codability review
Excluded at screening	2	Grade D unverifiable social-media claims
Included in preserved screening ledger	32	Retained rows before later retained-corpus maintenance
Reclassified out of codable retained corpus before public release	1	`2025-MLP-02` retained as contextual benchmark evidence only
New codable row added before the freeze	1	`2025-GEM-01`
Final public retained corpus	32	Matches `incident_registry_coded.csv` and the manuscript tables

A.8 Preserved Screened Ledger

The table below preserves the screened ledger that underlies the retained-corpus maintenance note above. It contains the 34 records advanced to incident-level screening in the preserved ledger state. The machine-readable `screening_ledger.csv` carries the release-stage dispositions: 31 `retained_in_registry_corpus` rows, 1 `retained_as_contextual_only` row, and 2 `excluded_at_screening` Grade D rows.

Log ID	Incident ID	Source type	Public identifier / access pathway	Decision	Exclusion reason
R001	2017-IG-01	Coroner/inquest	North London Coroner's Court inquest record (2022) documenting Instagram amplification pathway	Included	—
R002	2017-PIN-01	Coroner/inquest	Same inquest record; Pinterest recommendation-email pathway (retained as separate pathway record)	Included	—
R003	2023-CHA-01	Journalism	La Libre Belgique (28 Mar 2023) and Vice/Motherboard (30 Mar 2023) reporting with described log excerpts	Included	—
R004	2023-CAI-01	Court filing	D. Colo. No. 1:25-cv-02907 (filed 15 Sep 2025) (complaint); CourtListener docket available	Included	—
R005	2024-CAI-02	Court order	M.D. Fla. No. 6:24-cv-01903 (Order, 21 May 2025); CourtListener docket available	Included	—
R006	2024-CAI-01	Court filings	E.D. Tex. No. 2:24-cv-01014 (filed 9 Dec 2024) (complaint; arbitration order Doc. 59, 23 Apr 2025); CourtListener docket available	Included	—
R007	2025-GPT-01	Court filings	JCCP No. 5431 / Case No. CGC-25-628528 (ChatGPT Product Liability Cases; filings dated 26 Dec 2025) with reproduced excerpts	Included	—

Table 15 continued

Log ID	Incident ID	Source type	Public identifier / access pathway	Decision	Exclusion reason
R008	2024-GPT-01	Court filing	Complaint excerpt without docket citation in the v1.1 source file (retained as Grade C excerpt)	Included	—
R009	2025-MAI-01	Evaluation report	Common Sense Media Meta AI Risk Assessment (Aug 2025), systematic testing with ages 13–17	Included	—
R010	2025-NOM-01	Journalism	MIT Technology Review reporting (Jan–Apr 2025) with user screenshots regarding Nomi AI outputs	Included	—
R011	2025-REP-01	Complaint/filing	FTC complaint filing (Tech Justice Law Project, Young People's Alliance, Encode) (2025) alleging dependency and crisis-detection gaps	Included	—
R012	2025-THR-01	Peer-reviewed study	Moore et al. (FAccT 2025) evaluation of therapy chatbots (systematic red-team)	Included	—
R013	2025-MLP-03	Red-team study	Schoene & Canca (2025), arXiv:2507.02990 (jailbreaking in self-harm contexts)	Included	—
R014	2025-MLP-02	Peer-reviewed study	McBain et al. (2025), Psychiatric Services (LLM alignment with expert clinicians)	Included	—
R015	2025-MHB-01	Peer-reviewed study	Pichowicz et al. (2025), Scientific Reports 15:31652 (mental-health chatbot benchmark)	Included	—
R016	2025-MLP-01	Government hearing	U.S. Senate Judiciary Subcommittee hearing transcript: Examining the Harm of AI Chatbots (16 Sep 2025)	Included	—
R017	2024-GPT-02	Journalism	The New York Times (18 Aug 2025) first-person account reproducing selected chat excerpts (paywalled access possible)	Included	—
R018	2025-ACC-01	Journalism	ABC News / triple j Hack interview (Aug 2025) (pseudonymized account; no transcript excerpts published)	Included	—
R019	2025-GPT-05	Journalism	ABC News / triple j Hack interview (Aug 2025) (pseudonymized account; no transcript excerpts published)	Included	—
R020	2024-CAI-03	Court filing	N.D.N.Y. No. 1:25-cv-01295 (filed 16 Sep 2025) (complaint)	Included	—
R021	2025-CAI-01	Court filing	D. Colo. No. 1:25-cv-02906 (filed 15 Sep 2025) (complaint); CourtListener docket available	Included	—
R022	2025-GPT-10	Court filing	Cal. Super. Ct. S.F. City & Cty. No. CGC-25-630808 (filed 6 Nov 2025) (complaint)	Included	—

Table 15 continued

Log ID	Incident ID	Source type	Public identifier / access pathway	Decision	Exclusion reason
R023	2025-GPT-02	Party statement	Social Media Victims Law Center press release (6 Nov 2025) (party statement; no reproduced excerpts)	Included	—
R024	2025-GPT-03	Party statement	Social Media Victims Law Center press release (6 Nov 2025) (party statement; no reproduced excerpts)	Included	—
R025	2025-GPT-06	Court filing	Cal. Super. Ct. S.F. City & Cty. (complaint dated 11 Dec 2025; case number not stated in the copy used)	Included	—
R026	2025-GPT-12	Court filing	Cal. Super. Ct. L.A. County No. 25STCV32382 (filed 6 Nov 2025) (complaint)	Included	—
R027	2025-GPT-08	Court filing	Cal. Super. Ct. S.F. City & Cty. No. CGC-25-630809 (filed 6 Nov 2025) (complaint)	Included	—
R028	2025-GPT-09	Court filing	Cal. Super. Ct. S.F. City & Cty. No. CGC-25-630811 (filed 6 Nov 2025) (complaint + exhibits; screenshot-verified messages)	Included	—
R029	2025-GPT-11	Court filing	Cal. Super. Ct. L.A. County No. 25STCV32383 (filed 6 Nov 2025) (complaint; screenshot-verified messages)	Included	—
R030	2025-GPT-07	Court filing	Cal. Super. Ct. L.A. County No. 25STCV32386 (filed 6 Nov 2025) (amended complaint)	Included	—
R031	2025-GPT-04	Court filing	Cal. Super. Ct. L.A. County No. 25STCV32379 (filed 6 Nov 2025) (complaint)	Included	—
R032	2025-GPT-13	Court filing	Los Angeles County Superior Court (complaint filed Jan 2026; case number not stated in the copy used)	Included	—
R033	2025-012	Social media claim	Unverifiable social media claim (no attributable primary documentation recovered)	Excluded	Unverifiable; fails minimum reliability (Grade D)
R034	2025-013	Social media claim	Unverifiable social media claim (no attributable primary documentation recovered)	Excluded	Unverifiable; fails minimum reliability (Grade D)

A.9 Retained-Corpus Maintenance Update

The preserved screened ledger above is not itself the final public retained corpus. One older included contextual benchmark row was removed from the codable incident corpus, and one new A2 complaint row was added before the March 4, 2026 freeze.

Change type	Record	Effect on the public retained corpus	Reason
Reclassified out of codable retained corpus	`2025-MLP-02`	Removed from the 32-row incident retained corpus	Retained as contextual benchmark evidence rather than a codable incident row
New codable incident added	`2025-GEM-01`	Added to the 32-row incident retained corpus	New A2 complaint record, filed March 4, 2026, before the search cutoff

A.10 Policy-Corpus Assembly Rule

The parallel policy corpus was assembled separately from the incident ledger. A document entered the policy corpus only if it met all of the following conditions:

1. It was issued publicly by one of the five provider groups in the crosswalk.
2. It belonged to a codable governance genre: preparedness framework, responsible-scaling policy, system card, model card, model spec, usage policy, or comparable official safety/governance post.
3. It was current through the March 4, 2026 cutoff.
4. It was codable under `policy_registry_schema.csv` and not merely a superseded draft or non-governance marketing page.
5. The provider either appeared in the incident corpus or functioned as a major frontier-model governance comparator with a comparable public document set.

Companion-app vendors without a comparable public governance-document set by the freeze date remain out of scope for this crosswalk even when they appear in the incident registry.

B Definitions and Coding Boundaries

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Public release date: 2026-03-25

B.1 Core Objects

Term	Working definition
`AI-self-harm intersection`	The broad domain in which an AI system and a self-harm pathway co-occur in a documentable way.
`incident`	The underlying qualifying documented event.
`incident record`	The codable row tied to a specific harmed individual x platform/pathway exposure or a specific evaluation/demonstration record.
`pathway duplicate`	A retained second incident record for a distinct platform or pathway exposure involving the same harmed person.
`self-harm pathway`	A system-mediated sequence of exposure, reinforcement, instruction, escalation failure, or longitudinal dynamics that plausibly increases self-harm risk.
`trajectory-structured harm`	Harm that accrues across turns or sessions and depends on pathway dynamics, product affordances, or cross-session accumulation rather than a single output alone.
`human_harm_incident`	An incident record with a documented person-level death, injury, or harm-exposure outcome.
`evaluation_or_demonstration`	A benchmark, red-team, hearing-demonstration, or app-evaluation record retained because it documents unsafe behavior or governance-relevant failure signatures without a person-level injury/death event.

B.2 Outcome Labels

Use the canonical incident labels:

Outcome label	Definition
`DEATH`	A fatality is publicly reported.
`INJURY`	A non-fatal physical injury or self-harm injury is publicly reported.
`HARM_EXPOSURE`	Documented harm exposure, dependency, manipulation, or severe risk without a coded injury/death outcome.
`UNSAFE_OUTPUT`	Benchmark, red-team, hearing-demonstration, or app-evaluation evidence of unsafe output without a coded person-level injury/death event.

B.3 Reliability Grades and Evidentiary Levels

Reliability grades describe the traceability of the incident's public evidence anchor, not adjudicated merit.

Grade / level	Definition
`A1`	High-traceability primary-source record, such as a coroner finding, court order with reproduced excerpts, formal hearing transcript, or benchmark/evaluation report with directly inspectable methods and results.
`A2`	Formal allegation or complaint record that remains unadjudicated.
`B`	Investigative or first-person journalism with meaningful evidentiary detail and inspectable excerpts or sourcing, but without the traceability of A1/A2.
`C`	Lower-traceability secondary reporting or party statements where the incident is still codable but evidentiary limits are material.
`Level 1`	Documented incident record or benchmark/evaluation record anchored in inspectable materials.
`Level 2`	Formal allegation/mechanism record that remains unadjudicated.
`Level 3`	Paper-level governance inference; not assigned to seeded incident rows.

B.3.1 Explicit B/C Boundary

- `B` requires inspectable evidentiary detail beyond bare assertion, such as reproduced excerpts, screenshots, or attributable first-person / investigative sourcing that allows a reviewer to trace why the row is codable.
- `C` is used when the public record still supports row-level coding, but the available material is materially thinner: party statements, summary reporting, or limited excerpts without enough independent inspection to justify `B`.

B.4 Taxonomy Key

The registry uses a six-category, 13-subcode, non-mutually-exclusive taxonomy. These codes identify loci of evaluative intervention rather than mutually exclusive harm types.

Category	Subcode	Working definition
GEN	GEN-E	Encouragement: validates or fails to discourage suicidal intent.
GEN	GEN-C	Coaching: provides actionable self-harm instructions.
GEN	GEN-M	Method provision: addresses lethality or means access.
DET	DET-FN	Detection non-escalation: detects risk but does not transition to effective interruption, referral, or review.
LONG	LONG-DEP	Dependency / erosion of offline protective factors over time.
LONG	LONG-DEG	Worsening safety behavior across repeated turns or sessions.
LONG	LONG-DEL	Belief reinforcement inconsistent with shared reality across turns or sessions.
LONG	LONG-MEM	Memory resurfacing across sessions.
JB	JB-RP	Role-play framing bypass.
JB	JB-MT	Multi-turn bypass.
JB	JB-AC	Academic or research framing bypass.
RTI	RTI-CO	Fixed-belief non-amplification failure in a discrete exchange.
REC/MOD	REC/MOD-AMP	Recommendation or platform-architecture amplification.

B.5 Boundary Rules

B.5.1 RTI-CO Versus LONG-DEL

- Code RTI-CO when the record documents a discrete fixed-belief non-amplification failure in one focal exchange.
- Code LONG-DEL when the record documents repeated or temporally extended reinforcement of that frame across turns or sessions.
- Code both when the public record shows a focal affirming exchange embedded in a longer reinforcement trajectory.

B.5.2 LONG-DEP Versus High Engagement

Do not code LONG-DEP for mere frequency of use. The record should show dependency, emotional exclusivity, erosion of offline supports, or a closely related attachment pattern.

B.5.3 DET-FN

DET-FN is not a general criticism of safety quality. It is reserved for cases where the system detects or is presented with acute risk cues but does not meaningfully interrupt, route, or escalate.

B.6 Derived Analysis View

incident_registry_analysis_view.csv is an analysis-only companion keyed by incident_record_id. It does not alter the canonical 20-field incident schema.

B.6.1 Legacy identifier note

- incident_record_id is the row-level identifier used for the analysis companion.
- incident_id is retained in the same file as a backward-compatible legacy record identifier because downstream study materials already use that label.
- In the current public release, incident_record_id and incident_id are identical string values.

B.6.2 record_type

- human_harm_incident for person-level death, injury, or harm-exposure rows.
- evaluation_or_demonstration for rows where user_type = test.

B.6.3 evaluation_subtype

Use the following mapping for the six evaluation/demonstration rows:

Incident ID	Evaluation subtype (evaluation_subtype)
2025-MAI-01	App evaluation (app_evaluation)
2025-MHB-01	Benchmark study (benchmark_study)
2025-MLP-01	Hearing demonstration (hearing_demonstration)
2025-MLP-03	Red-team exercise (red_team)
2025-NOM-01	App evaluation (app_evaluation)
2025-THR-01	Benchmark study (benchmark_study)

B.6.4 trajectory_structured_strict

- yes when one or more adjudicated LONG-* subcodes are present.
- no otherwise.

This field is fully mechanical: it is derived from the canonical taxonomy coding only, and it is the manuscript's primary trajectory indicator.

B.6.5 trajectory_basis_strict

- Use semicolon-separated LONG-* subcodes when trajectory_structured_strict = yes.
- Use none when trajectory_structured_strict = no.

B.6.6 trajectory_structured_broad

- yes when the record satisfies the strict rule or the public record otherwise documents a cross-turn or cross-session pathway strongly enough to satisfy the prespecified appendix rule.
- no when only event-level evidence is supported.
- indeterminate when the public materials describe an extended pathway but remain too thin for a conservative broad-route yes.

This field is retained as a descriptive sensitivity check rather than the headline prevalence measure.

B.6.7 trajectory_basis_broad

- Use semicolon-separated LONG-* subcodes when trajectory_structured_broad = yes because of adjudicated longitudinal codes.
- Use documented_multi_turn_pathway_insufficient_for_subcode when trajectory_structured_b = indeterminate.
- Use none when trajectory_structured_broad = no.

B.6.8 Legacy trajectory alias

- trajectory_structured_flag is retained as a backward-compatible alias of trajectory_structured_broad.
- trajectory_basis is retained as a backward-compatible alias of trajectory_basis_broad.

B.6.9 One-page decision tree

1. Does the adjudicated registry row contain one or more `LONG-*` subcodes? If yes, set `trajectory_structured_strict = yes`, `trajectory_structured_broad = yes`, and carry the `LONG-*` codes into both basis fields.
2. If no `LONG-*` code is present, does the public record still document repeated turns, repeated sessions, retained memory, or another clearly cumulative interaction pathway? If no, set `trajectory_structured_strict = no` and `trajectory_structured_broad = no`.
3. If the public record does document a repeated or cumulative pathway without adjudicated `LONG-*` support, ask whether the evidence is specific enough to support a conservative broad-route longitudinal designation. If yes, set `trajectory_structured_broad = yes`. If no, set `trajectory_structured_broad = indeterminate`.

B.6.10 Worked examples

Incident ID	Strict trajectory flag (`trajectory_structured_strict`)	Broad trajectory flag (`trajectory_structured_broad`)	Rationale
`2025-GEM-01`	`yes`	`yes`	Adjudicated `LONG-DEL`, `LONG-DEP`, and `LONG-MEM` make the strict route mechanical.
`2025-GPT-04`	`yes`	`yes`	`LONG-DEL` is present, so both flags resolve `yes`.
`2024-GPT-02`	`no`	`indeterminate`	The public account suggests a repeated pathway, but the released evidence remains too thin for conservative `LONG-*` assignment.
`2025-MHB-01`	`no`	`no`	The benchmark shows unsafe outputs without a retained multi-turn trajectory object at the released row level.
`2025-MAI-01`	`yes`	`yes`	`LONG-MEM` is explicitly coded from the evaluation evidence.

B.6.11 Cluster Fields

- `person_cluster_id` groups explicit pathway duplicates or clearly identical harmed-person rows for analytic deduplication.
- `provider_cluster_id` groups rows by provider or platform family.
- `case_cluster_id` groups rows by the same case family, inquest, hearing package, or benchmark study.
- `pathway_duplicate_group_id` groups rows that document the same underlying harmed-person pathway but are intentionally retained as separate platform/pathway records.

B.7 Study Architecture

Layer	Unit	What it contributes
Registry layer	32 incident records	Incident metadata, outcome category, evidence grade, and evidence anchor.
Taxonomy layer	13 multi-label subcodes	Governance-relevant failure loci across `GEN`, `DET`, `LONG`, `JB`, `RTI`, and `REC/MOD`.
Analysis layer	Derived analysis view	`incident_record_id`, strict and broad trajectory fields, and cluster-aware sensitivity fields.
Policy crosswalk layer	16 public governance documents	Public placement of self-harm in preparedness, product-safety, and trust-and-safety documents.

B.8 Policy Crosswalk Glossary

Shorthand	Meaning
`Class IV`	Reactive incident-domain treatment, such as post-incident patches or case-specific remediation.
`native`	A policy document names self-harm as an evaluative object and pairs it with a usable evaluation procedure.
`proxy_represented`	A policy document addresses self-harm only through adjacent categories such as dangerous content or user well-being.
`operationally_underspecified`	A policy document names the domain but does not provide a usable prospective evaluation procedure.
`Absent / not represented`	Used as the absence label for both `evaluative_object_status` and `risk_domain_status`; self-harm is not present in codable form in the document.
`Ambiguous / insufficiently classifiable`	Self-harm is mentioned but the public text does not support a defensible `Class I` to `Class IV` placement.
Recorded governance tier (`risk_domain_status`)	The highest governance tier actually evidenced for self-harm in the document, coded as `Class I`, `Class II`, `Class III`, `Class IV`, `Absent / not represented`, or `Ambiguous / insufficiently classifiable`.

Shorthand	Meaning
Coder confidence (`coder_confidence`)	Confidence based on document clarity, not on agreement with the organization's framing.
Document genre (`document_genre`)	Broader analytic grouping used for within-genre policy comparison before provider-level rollups.
All self-harm layers present (`all_self_harm_layers_present`)	Semicolon-separated list of every governance layer in the document that explicitly carries self-harm treatment.

B.9 One-Page Glossary for Frequent Shorthand

Shorthand	Meaning
`full retained corpus`	All 32 incident rows in `incident_registry_coded.csv`.
`human-harm subset`	The 27 rows with person-level death, injury, or harm-exposure outcomes.
`A1/A2 primary robustness subset`	The 20 human-harm rows with higher-traceability A1 or A2 evidence.
`A1/A2/B secondary sensitivity subset`	The 22 human-harm rows with A1, A2, or B evidence.
`C-excluded mechanism sensitivity`	Human-harm rows after excluding C-grade records that still carry mechanism subcodes.

C Reliability Appendix

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Public release date: 2026-03-25

C.1 Source Basis

This appendix now distinguishes between two reliability components:

1. the retained 8-incident independent-rater summary audit already archived in earlier study materials; and
2. the new full-corpus audit scaffold archived in `incident_taxonomy_rater_matrix.csv` and `incident_taxonomy_adjudication_log.csv`.

Historical raw second-rater exports for the retained 8-incident audit are still **not** publicly available. The new full-corpus files therefore expose the primary-coder side of the audit matrix and the forward adjudication surface, but they should not be described as a completed independent second-rater re-audit.

C.2 Retained 8-Incident Audit Summary

C.2.1 Recoverable audited incident set

The retained audit summary covers the following incident records:

- `2017-IG-01`
- `2023-CAI-01`
- `2024-CAI-02`
- `2025-GEM-01`
- `2025-GPT-01`
- `2025-MAI-01`
- `2025-MLP-01`
- `2025-REP-01`

C.2.2 Coverage note

- The recoverable audited set covers deaths, harm exposures, and unsafe-output rows.
- The recoverable audited set does **not** include an injury row.
- The recoverable audited set is concentrated in `A1` and `A2` materials; it does not provide a raw archived audit surface for the most interpretation-sensitive `B` and `C` rows.

The retained 8-incident table is an archived historical agreement summary preserved from an earlier coding state. It should not be read as a row-level parity check or independent validation of the current 2.0 released taxonomy assignments, which have since been updated. Accordingly, the pooled and per-subcode agreement statistics below are historical audit artifacts, not current-release subcode validation metrics.

C.2.3 Retained pooled agreement summary

Metric	Value
Incidents audited	8 of 32 (25.0%)
Total decisions (`incident x subcode`)	104
Both raters positive	23
Both raters negative	73
Rater 2 only	8
Rater 1 only	0
Raw agreement	92.3%
Pooled Cohen's kappa	0.80
Approximate 95% CI for pooled kappa	0.67 to 0.93

The confidence interval is the asymptotic interval implied by the retained pooled `2 x 2` decision table (`23 / 0 / 8 / 73`), not a bootstrap interval from raw coder-level exports.

C.2.4 Retained per-subcode summary

Subcode	Prevalence (either rater)	Raw agreement	Kappa	Positive agreement	PP	NN	R2 only	R1 only
GEN-E	4/8	8/8	1.00	1.00	4	4	0	0
GEN-C	3/8	7/8	0.71	0.80	2	5	1	0
GEN-M	2/8	6/8	0.00	0.00	0	6	2	0
DET-FN	4/8	7/8	0.75	0.86	3	4	1	0
LONG-DEP	5/8	8/8	1.00	1.00	5	3	0	0
LONG-DEG	1/8	7/8	0.00	0.00	0	7	1	0
LONG-DEL	3/8	8/8	1.00	1.00	3	5	0	0
LONG-MEM	1/8	7/8	0.00	0.00	0	7	1	0
JB-RP	1/8	8/8	1.00	1.00	1	7	0	0
JB-MT	1/8	7/8	0.00	0.00	0	7	1	0
JB-AC	1/8	8/8	1.00	1.00	1	7	0	0
RTI-CO	5/8	7/8	0.75	0.89	4	3	1	0
REC/MOD-AMP	1/8	8/8	1.00	1.00	1	7	0	0

For sparse subcodes, the agreement counts and positive agreement are more informative than kappa alone.

C.3 Full-Corpus Audit Scaffold

Two new files now define the forward reliability surface:

- `downloads/incidents/incident_taxonomy_rater_matrix.csv`
- `downloads/incidents/incident_taxonomy_adjudication_log.csv`

C.3.1 What the scaffold contains

- One row per `incident_record_id x review_item` combination across the full 32-record retained registry corpus.
- `review_family = taxonomy_subcode` for the 13 multi-label taxonomy decisions.
- Additional review rows for `outcome_category`, `reliability_grade`, `eligibility_retention_status`, and `trajectory_structured_broad`.
- `evidence_stratum` populated as `A1`, `A2`, `B`, or `C` for planned stratified reporting.
- `blinding_requirement = blind_to_rater_1_and_expected_outcomes` for every row.
- `rater_1` populated from the current released coding.

- Blank `rater_2` and `adjudicated_value` columns reserved for a future independent second-rater completion pass.
- A status column marking every row as `awaiting_independent_second_rater`.

C.3.2 What the scaffold does not contain

- It does not reconstruct historical second-rater decisions that were not archived.
- It does not justify stronger claims than the retained 8-incident summary audit supports.
- It does not convert the current release into a completed full-corpus independent-rater package.

C.3.3 Intended reporting strata for a completed pass

If a true second-rater pass is later completed, report agreement separately for:

- `A1/A2`
- `B`
- `C`

Do not collapse `B` and `C` into a single interpretive-risk stratum.

C.3.4 Planned review-surface rules

- `outcome_category` and `reliability_grade` should be coded independently from the released CSV.
- `eligibility_retention_status` should distinguish `retained_in_registry_corpus` from any future contextual-only or excluded rows if the corpus changes.
- `trajectory_structured_broad` requires independent review because it is not purely mechanical.
- `trajectory_structured_strict` does not require a second coder because it is mechanically derivable from adjudicated `LONG-*` values.

C.4 Interpretation-Seams Still Worth Monitoring

The retained audit materials and the current codebook continue to identify three seams that deserve attention in any future completed re-audit:

1. `GEN-E` versus no code when a response is ambiguous between encouragement and neutral acknowledgment.
2. `LONG-DEL` versus `RTI-CO`, where the practical distinction is temporal: focal exchange versus repeated reinforcement trajectory.
3. `LONG-DEP` versus high engagement without a documented dependency or exclusivity signal.

C.5 What This Appendix Supports

This appendix supports a narrower claim than the earlier wording: the current taxonomy is operationalized and partially reliability-tested, but the public materials do not yet contain a completed full-corpus independent second-rater archive. The retained 8-incident summary audit remains informative; the new scaffold makes the next audit pass forward-completable and file-level auditable.

D Policy Reliability Appendix

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Public release date: 2026-03-25

D.1 Purpose

This appendix defines the auditable reliability surface for the 16-document policy crosswalk and records what is and is not yet archived in the public release.

D.2 Archived Files

- `downloads/policy/policy_registry_second_rater.csv`
- `downloads/policy/policy_registry_adjudication_log.csv`
- `downloads/policy/policy_registry_coded.csv`
- `downloads/policy/policy_registry_schema.csv`
- `downloads/policy/policy_registry_coding_manual.md`

D.3 Current Status

The policy crosswalk remains a single-coder descriptive pass in this release. The new files add the row-level second-rater scaffold and adjudication surface, but they do **not** contain completed independent second-rater decisions.

D.4 Independent-Review Surface

The intended non-derived analytic review surface is:

- `document_genre`
- `self_harm_named`
- `self_harm_domain_form`
- `risk_domain_status`
- `first_class_equivalence`
- `named_frontier_risk_domains`
- `self_harm_comparability_note`
- `threshold_mechanism`
- `deployment_consequence`
- `external_review_requirement`
- `specialized_red_teaming`
- `monitoring_obligation`
- `incident_response_layer`
- `all_self_harm_layers_present`
- `governance_fragmentation`
- `temporal_model`
- `multi_turn_evaluation`
- `memory_personalization_attention`
- `dependency_attention`
- `cross_session_accumulation`
- `mechanism_coverage`
- `evaluative_object_status`
- `epistemic_risk_attention`
- `architecture_level_attention`
- `detection_to_action_coupling`

Derived gap fields should be regenerated only after those base fields are independently reviewed and adjudicated.

D.5 Scaffold Design

D.5.1 `policy_registry_second_rater.csv`

- One row per policy document.
- `record_id` populated for all 16 documents.
- `document_genre` and all analytic coding fields left blank pending an actual independent second-rater pass.
- `audit_status = awaiting_independent_second_rater` for every row.

D.5.2 `policy_registry_adjudication_log.csv`

- Reserved for field-level disagreements.
- Should be populated only after a completed second-rater pass exists.

D.6 Recommended Completion Rule

When a true second-rater pass is available:

1. Complete `policy_registry_second_rater.csv` independently from the canonical coded CSV.
2. Compare coder 1 versus coder 2 on the non-derived analytic fields only.
3. Log every disagreement in `policy_registry_adjudication_log.csv`.
4. Recompute derived gap fields from the adjudicated base fields.
5. Report exact agreement and Cohen's kappa for single-choice nominal fields.
6. Report exact-set agreement and label-level positive agreement for `mechanism_coverage`.
7. Report results within `document_genre` before any provider-level rollup.

D.7 What This Appendix Supports

This appendix supports two bounded claims:

- the public materials now provide a file-level, forward-completable reliability surface for the policy crosswalk; and
- the policy crosswalk should still be described as a single-coder descriptive audit until the scaffold is actually completed by an independent second rater.

E Policy Crosswalk Appendix

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Public release date: 2026-03-25

E.1 Purpose

This appendix makes the 16-document policy crosswalk directly auditable from the public study materials. It states the corpus-selection rule, summarizes the public-document field distributions, lists the fixed fields intended for independent review, and exposes the row-level coding ledger derived from `policy_registry_coded.csv`. It should be read together with `policy_reliability_appendix.md`, which provides the forward second-rater scaffold, and `policy_source_index.csv`, which records public source-capture metadata for the retained policy corpus.

E.2 Corpus Selection Rule

A document enters the public policy corpus only if it satisfies all of the following conditions:

1. It is publicly issued by one of the five selected provider groups represented in the crosswalk.
2. It belongs to a codable governance genre: preparedness framework, responsible-scaling policy, system card, model card, model spec, usage policy, or comparable official safety/governance post.
3. It is current through the March 4, 2026 search cutoff and policy corpus freeze.
4. It is codable under `policy_registry_schema.csv` using the highest-evidenced-tier rule.
5. It is not a superseded draft, marketing page, or non-codable announcement.
6. The provider either appears in the incident registry or functions as a major frontier-model governance comparator with a comparable public-document set by the freeze date.

The resulting corpus contains 16 documents across Anthropic, Google / Google DeepMind, Meta, OpenAI, and xAI. Companion-app vendors without a comparable public governance-document set by the freeze date remain out of scope even when they appear in the incident registry.

E.3 Provider-Group Freeze

Google and Google DeepMind are treated as one provider group for corpus selection and provider-level rollups, while their public documents retain separate issuing-organization labels at the row level.

Provider group	Inclusion basis at the March 4, 2026 freeze
Anthropic	Included as a frontier-model provider with a public set spanning responsible-scaling and product-safety / governance-post genres by the freeze.

Provider group	Inclusion basis at the March 4, 2026 freeze
Google / Google DeepMind	Included because Gemini-related materials are both incident-relevant and part of the frontier-governance comparison set, with codable documents issued under both Google and Google DeepMind labels by the freeze.
Meta	Included because Meta appears in the incident record and also had codable public governance documents in the predefined genres by the freeze.
OpenAI	Included because OpenAI appears repeatedly in the incident record and had both preparedness and product-safety documents publicly available by the freeze.
xAI	Included as a major frontier-model governance comparator with codable frontier and product-safety documents publicly available by the freeze.

E.4 Corpus Composition

E.4.1 By organization

Organization	Documents
Anthropic	5
Google	1
Google DeepMind	2
Meta	2
OpenAI	4
xAI	2

E.4.2 By document genre

Document genre	Documents
Frontier or scaling (`frontier_or_scaling`)	5
Product-safety artifact (`product_safety_artifact`)	5
Policy or governance post (`policy_or_governance_post`)	6

E.4.3 By genre group

Genre group	Documents	Tier profile	Native documents	Documents with any threshold mechanism	Documents with any multi-turn evaluation
Frontier or scaling (`frontier_or_scaling`)	5	Absent / not represented in all 5 documents	0	0	0
Product-safety artifact (`product_safety_artifact`)	5	Class II in 3 documents; Class III in 2	1	2	2
Policy or governance post (`policy_or_governance_post`)	6	Class II in 2 documents; Class III in 3; Absent / not represented in 1	1	0	3

E.4.4 High-level field summary

Field	Distribution
Self-harm named (`self_harm_named`)	Explicit in 10 documents; absent in 6
Self-harm domain form (`self_harm_domain_form`)	Subdomain in 6 documents; proxy domain in 2; native domain in 2; absent in 6
Recorded governance tier (`risk_domain_status`)	Class II in 5 documents; Class III in 5; absent / not represented in 6
Evaluative-object status (`evaluative_object_status`)	Native in 2 documents; proxy represented in 2; operationally underspecified in 6; ontologically absent in 6
First-class equivalence (`first_class_equivalence`)	No in all 16 documents
Threshold mechanism (`threshold_mechanism`)	Partial in 2 documents; none in 14
Deployment consequence (`deployment_consequence`)	Partial in 3 documents; none in 13
Multi-turn evaluation (`multi_turn_evaluation`)	Explicit in 3 documents; partial in 2; none in 11
All self-harm layers present (`all_self_harm_layers_present`)	Product safety in 8 documents; trust and safety in 2; absent in 6

E.5 Fixed Fields for Independent Review

The manuscript discusses the following fields as the minimum independent-review surface for the policy crosswalk:

- `document_genre`
- `self_harm_named`
- `self_harm_domain_form`
- `risk_domain_status`
- `evaluative_object_status`
- `first_class_equivalence`
- `threshold_mechanism`
- `deployment_consequence`
- `multi_turn_evaluation`

E.5.1 Public-archive status of independent review

The policy crosswalk remains a single-coder descriptive pass. The public materials include `policy_registry_second_rater.csv` and `policy_registry_adjudication_log.csv` as forward-completable audit files, but they do **not** yet include a completed independent second-rater review for all 16 documents. This appendix therefore exposes the row-level evidence, excerpts, and field values needed for such a review, but it should not be described as a completed independent adjudication table.

E.6 Row-Level Public-Document Ledger

The table below is generated from the current coded CSV and is the authoritative row-level audit surface for the manuscript. Two retained Meta canonical URLs on `ai.meta.com` required login during the March 25, 2026 release audit; the study materials preserve their canonical URLs together with source-capture metadata and archived excerpt provenance.

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
ANTHROPIC-2025-01	Anthropic	How people use Claude for support, advice, and companionship (blog_post; policy_or_governance_post)	2025-06-27	https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship	named=explicit;form=proxy_domain;tier=ClassIII;genre=policy_or_governance_post;layers=product_safety;object=proxy_represented;first_class=no;threshold=none;deploy=none;multi_turn=partial	when it does, it's typically for safety reasons . . . refusing to provide dangerous weight loss advice or support self-harm.	moderate
ANTHROPIC-2025-02	Anthropic	Building safeguards for Claude (blog_post; policy_or_governance_post)	2025-08-12	https://www.anthropic.com/news/building-safeguards-for-claude	named=explicit;form=subdomain;tier=ClassII;genre=policy_or_governance_post;layers=product_safety;object=operationally_underspecified;first_class=no;threshold=none;deploy=partial;multi_turn=explicit	We assess Claude's adherence to our Usage Policy on topics like child exploitation or self-harm . . . including . . . extended multi-turn conversations.	high
ANTHROPIC-2025-03	Anthropic	Protecting the well-being of our users (blog_post; policy_or_governance_post)	2025-12-18	https://www.anthropic.com/news/protecting-well-being-of-users	named=explicit;form=native_domain;tier=ClassII;genre=policy_or_governance_post;layers=product_safety;object=native;first_class=no;threshold=none;deploy=none;multi_turn=explicit	We focus on two areas: how Claude handles conversations about suicide and self-harm . . . we use a combination of model training and product interventions.	high

Table 15 continued

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
ANTHROPIC-2025-04	Anthropic	Sharing our compliance framework for California's Transparency in Frontier AI Act (`blog_post`; `policy_or_governance_post`)	2025-12-19	https://www.anthropic.com/news/compliance-framework-SB53	`named=absent;form=absent;tier=Absent/notrepresented;genre=policy_or_governance_post;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	Our FCF describes how we assess and mitigate cyber offense, chemical, biological, radiological, and nuclear threats . . . as well as the risks of AI sabotage and loss of control.	high
ANTHROPIC-2026-01	Anthropic	Anthropic's Responsible Scaling Policy: Version 3.0 (`responsible_scaling_policy`; `frontier_or_scaling`)	2026-02-24	https://www.anthropic.com/news/responsible-scaling-policy-v3	`named=absent;form=absent;tier=Absent/notrepresented;genre=frontier_or_scaling;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	Non-novel chemical/biological weapons production . . . High-stakes sabotage opportunities . . . Automated R&D in key domains.	high
GOOGLE-2024-01	Google	Generative AI Prohibited Use Policy (`usage_policy`; `policy_or_governance_post`)	2024-12-17	https://policies.google.com/terms/generative-ai/use-policy	`named=explicit;form=subdomain;tier=Class III;genre=policy_or_governance_post;layers=trust_and_safety;object=operationally_underspecified;first_class=no;threshold=none;deploy=none;multi_turn=none`	Facilitates self-harm.	high

Table 15 continued

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
GOOGLEDEEPMIND-2025-01	Google DeepMind	Frontier Safety Framework Report - Gemini 3 Pro (November, 2025) v2 (`frontier_safety_framework`; `frontier_or_scaling`)	2025-11	https://deepmind.google/models/fsf-reports/gemini-3-pro/	`named=absent;form=absent;tier=Absent/notrepresented;genre=frontier_or_scaling;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	It currently covers four risk domains . . . CBRN . . . , cybersecurity, machine learning R&D, and harmful manipulation, and also includes . . . misalignment risk.	high
GOOGLEDEEPMIND-2025-02	Google DeepMind	Gemini 3 Pro - Model Card (`model_card`; `product_safety_artifact`)	2025-12	https://deepmind.google/models/model-cards/gemini-3-pro	`named=explicit;form=proxy_domain;tier=ClassIII;genre=product_safety_artifact;layers=product_safety;object=proxy_represented;first_class=no;threshold=none;deploy=none;multi_turn=partial`	Dangerous content (e.g., promoting suicide, or instructing in activities that could cause real-world harm).	high
META-2023-01	Meta	Llama 2 Acceptable Use Policy (`usage_policy`; `policy_or_governance_post`)	2023-07-18	https://ai.meta.com/llama/use-policy/	`named=explicit;form=subdomain;tier=ClassIII;genre=policy_or_governance_post;layers=trust_and_safety;object=operationally_underspecified;first_class=no;threshold=none;deploy=none;multi_turn=none`	Self-harm or harm to others, including suicide, cutting, and eating disorders.	moderate

Table 15 continued

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
META-2025-01	Meta	Code World Model Preparedness Report (`preparedness_framework`; `frontier_or_scaling`)	2025-09-24	https://ai.meta.com/research/publications/code-world-model-preparedness-report/	`named=absent;form=absent;tier=Absent/not represented;genre=frontier_or_scaling;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	we conducted an automated assessment of CWM capabilities in two domains . . . namely Cybersecurity and Chemical & Biological risks.	high
OPENAI-2025-01	OpenAI	Preparedness Framework (Version 2) (`preparedness_framework`; `frontier_or_scaling`)	2025-04-15	https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf	`named=absent;form=absent;tier=Absent/not represented;genre=frontier_or_scaling;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	We currently focus this work on three areas of frontier capability . . . Biological and Chemical . . . Cybersecurity . . . AI Self-improvement capabilities.	high
OPENAI-2025-02	OpenAI	Addendum to GPT-5 System Card: Sensitive Conversations (`system_card`; `product_safety_artifact`)	2025-10-27	https://cdn.openai.com/pdf/3da476af-b937-47fb-9931-88a851620101/addendum-to-gpt-5-system-card-sensitive-conversations.pdf	`named=explicit;form=subdomain;tier=Class II;genre=product_safety_artifact;layers=product_safety;object=operationally_underspecified;first_class=no;threshold=partial;deploy=partial;multi_turn=none`	self-harm/intent . . . self-harm/instructions	high

Table 15 continued

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
OPENAI-2025-03	OpenAI	Model Spec (2025/12/18) (`model_spec`; `product_safety_artifact`)	2025-12-18	https://model-spec.openai.com/2025-12-18.html	`named=explicit;form=native_domain;tier=ClassII;genre=product_safety_artifact;layers=product_safety;object=operationally_underspecified;first_class=no;threshold=none;deploy=none;multi_turn=none`	The assistant must not encourage or enable self-harm . . . always advising that immediate help should be sought if the user is in imminent danger.	high
OPENAI-2026-01	OpenAI	GPT-5.3 Instant System Card (`system_card`; `product_safety_artifact`)	2026-03-03	https://deploymentsafety.openai.com/gpt-5-3-instant/gpt-5-3-instant.pdf	`named=explicit;form=subdomain;tier=ClassII;genre=product_safety_artifact;layers=product_safety;object=native;first_class=no;threshold=partial;deploy=partial;multi_turn=explicit`	we implemented dynamic multi-turn evaluations for mental health, emotional reliance, and self-harm that simulate extended conversations across these domains.	high
XAI-2025-01	xAI	Grok 4.1 Model Card (`model_card`; `product_safety_artifact`)	2025-11-17	https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf	`named=explicit;form=subdomain;tier=ClassIII;genre=product_safety_artifact;layers=product_safety;object=operationally_underspecified;first_class=no;threshold=none;deploy=none;multi_turn=none`	we employ input filters to reject specific classes of sensitive requests, such as those involving bioweapons, chemical weapons, self-harm, and child sexual abuse material.	high

Table 15 continued

Record ID	Organization	Document	Date	URL	Key coded fields	Supporting excerpt	Confidence
XAI-2025-02	xAI	xAI Frontier Artificial Intelligence Framework (`frontier_safety_framework`; `frontier_or_scaling`)	2025-12-30	https://data.x.ai/2025-12-31-xai-frontier-artificial-intelligence-framework.pdf	`named=absent;form=absent;tier=Absent/notrepresented;genre=frontier_or_scaling;layers=absent;object=Absent/notrepresented;first_class=no;threshold=none;deploy=none;multi_turn=none`	This FAIF discusses two major categories of AI risk - malicious use and loss of control.	high

E.7 Reading the Table

- `tier` is the coded `risk_domain_status`.
- `genre` is the coded `document_genre`, used for within-genre comparison before provider-level rollups.
- `layers` is the coded `all_self_harm_layers_present`.
- `object` is the coded `evaluative_object_status`.
- `first_class` indicates whether self-harm is treated on frontier-comparable terms.
- `threshold` and `deploy` report whether self-harm findings trigger thresholded review or deployment consequences in the public document itself.
- `multi_turn` reports whether the document contains explicit, partial, or absent longitudinal evaluation procedures.

E.8 What This Appendix Supports

This appendix supports cautious claims about what the **public documents themselves** operationalize. It does not support claims about undisclosed internal practice, unpublished red-team protocols, or organization-wide governance measures not evidenced in the document row being coded.

F Reference and Source Index

Version: 2.0
Search cutoff / incident corpus freeze / policy corpus freeze: 2026-03-04
Public release date: 2026-03-25

This appendix indexes the public evidence routes used in the registry and the public policy documents included in the crosswalk. Datasets, coding manuals, citation metadata, and provenance materials are available from the downloads page.

F.1 Incident Citation and Evidence-Route Index

`Archive status = present` means a local evidence file is bundled in the release. `Archive status = external_only` means the public route is an external primary source and no local mirror is bundled.

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2017-IG-01	Meta Instagram	DEATH	A1	/incidents/2017-IG-01_Coroner_Report.pdf	/evidence/incidents/2017-IG-01/2017-IG-01_Coroner_Report.pdf	present
2017-PIN-01	Pinterest	DEATH	A1	/incidents/2017-PIN-01_Coroner_Report.pdf	/evidence/incidents/2017-PIN-01/2017-PIN-01_Coroner_Report.pdf	present
2023-CAI-01	Character.AI	DEATH	A2	1:25-cv-02907; /incidents/2023-CAI-01.pdf	/evidence/incidents/2023-CAI-01/2023-CAI-01.pdf	present
2023-CHA-01	Chai Research (EleutherAI GPT-J fine-tuned)	DEATH	B	/incidents/2023-CHA-01_LaLibre.pdf; /incidents/2023-CHA-01_Vice.pdf	/evidence/incidents/2023-CHA-01/2023-CHA-01_LaLibre.pdf	present
2024-CAI-01	Character.AI	INJURY	A2	2:24-cv-01014; /incidents/2024-CAI-01.pdf; /incidents/2024-CAI-01_Order.pdf	/evidence/incidents/2024-CAI-01/2024-CAI-01.pdf	present
2024-CAI-02	Character.AI	DEATH	A1	6:24-cv-01903; /incidents/2024-CAI-02.pdf; /incidents/2024-CAI-02_Order.pdf	/evidence/incidents/2024-CAI-02/2024-CAI-02.pdf	present
2024-CAI-03	Character.AI	INJURY	A2	1:25-cv-01295; /incidents/2024-CAI-03.pdf	/evidence/incidents/2024-CAI-03/2024-CAI-03.pdf	present

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2024-GPT-01	OpenAI ChatGPT	HARM_EXPOSURE	C	Christopher 'Kirk' Shamblin and Alicia Shamblin, individually and as successors-in-interest to Decedent, Zane Shamblin v. OpenAI, Inc., et al., No. 25STCV32382 (Cal. Super. Ct., Los Angeles County, filed Nov. 6, 2025); `/incidents/2024-GPT-01.pdf`	`/evidence/incidents/2024-GPT-01/2024-GPT-01.pdf`	present
2024-GPT-02	OpenAI ChatGPT (GPT-4o, persona: "Harry")	DEATH	B	https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html	https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html	external_only
2025-ACC-01	AI companion chatbots (multiple; vendor unspecified)	HARM_EXPOSURE	C	`/incidents/2025-ACC-01_ABC.pdf`	`/evidence/incidents/2025-ACC-01/2025-ACC-01_ABC.pdf`	present
2025-CAI-01	Character.AI	HARM_EXPOSURE	A2	1:25-cv-02906; `/incidents/2025-CAI-01.pdf`	`/evidence/incidents/2025-CAI-01/2025-CAI-01.pdf`	present
2025-GEM-01	Google Gemini 2.5 Pro	DEATH	A2	5:26-cv-01849-VKD; `/incidents/2025-GEM-01.pdf`	`/evidence/incidents/2025-GEM-01/2025-GEM-01.pdf`	present

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-GPT-01	OpenAI ChatGPT (GPT-4o)	DEATH	A2	CGC-25-628528; `/incidents/2025-GPT-01.pdf`; `/incidents/2025-GPT-01_JCCP_Opposition.pdf`	`/evidence/incidents/2025-GPT-01/2025-GPT-01.pdf`	present
2025-GPT-02	OpenAI ChatGPT (GPT-4o)	HARM_EXPOSURE	C	https://techjusticelaw.org/2025/11/06/social-media-victims-law-center-and-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/	https://techjusticelaw.org/2025/11/06/social-media-victims-law-center-and-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/	external_only
2025-GPT-03	OpenAI ChatGPT (GPT-4o)	HARM_EXPOSURE	C	https://techjusticelaw.org/2025/11/06/social-media-victims-law-center-and-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/	https://techjusticelaw.org/2025/11/06/social-media-victims-law-center-and-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/	external_only
2025-GPT-04	OpenAI ChatGPT (GPT-4o)	DEATH	A2	25STCV32379; `/incidents/2025-GPT-04.pdf`	`/evidence/incidents/2025-GPT-04/2025-GPT-04.pdf`	present

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-GPT-05	OpenAI ChatGPT (model unspecified)	INJURY	C	/incidents/2025-GPT-05_ABC.pdf	/evidence/incidents/2025-GPT-05/2025-GPT-05_ABC.pdf	present
2025-GPT-06	OpenAI ChatGPT (GPT-4o; ChatGPT Plus)	DEATH	A2	CGC-25-631477; /incidents/2025-GPT-06.pdf	/evidence/incidents/2025-GPT-06/2025-GPT-06.pdf	present
2025-GPT-07	OpenAI ChatGPT (GPT-4o Plus)	HARM_EXPOSURE	A2	25STCV32386; /incidents/2025-GPT-07.pdf	/evidence/incidents/2025-GPT-07/2025-GPT-07.pdf	present
2025-GPT-08	OpenAI ChatGPT (GPT-4o)	HARM_EXPOSURE	A2	Karen Enneking, individually and as successor-in-interest to decedent Joshua Enneking v. OpenAI, Inc., et al., No. CGC-25-630809 (Cal. Super. Ct., San Francisco County, filed Nov. 6, 2025); /incidents/2025-GPT-08.pdf	/evidence/incidents/2025-GPT-08/2025-GPT-08.pdf	present
2025-GPT-09	OpenAI ChatGPT (GPT-4o)	INJURY	A2	CGC-25-630811; /incidents/2025-GPT-09.pdf	/evidence/incidents/2025-GPT-09/2025-GPT-09.pdf	present
2025-GPT-10	OpenAI ChatGPT (GPT-4o)	DEATH	A2	CGC-25-630808; /incidents/2025-GPT-10.pdf	/evidence/incidents/2025-GPT-10/2025-GPT-10.pdf	present
2025-GPT-11	OpenAI ChatGPT (GPT-4o)	INJURY	A2	25STCV32383; /incidents/2025-GPT-11.pdf	/evidence/incidents/2025-GPT-11/2025-GPT-11.pdf	present

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-GPT-12	OpenAI ChatGPT	DEATH	A2	Christopher 'Kirk' Shamblin and Alicia Shamblin, individually and as successors-in-interest to Decedent, Zane Shamblin v. OpenAI, Inc., et al., No. 25STCV32382 (Cal. Super. Ct., Los Angeles County, filed Nov. 6, 2025); https://techjusticelaw.org/2025/11/06/social-media-victims-law-center-and-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/; https://selfharm.ai/incidents/2025-GPT-12.pdf; `/incidents/2025-GPT-12.pdf`	`/evidence/incidents/2025-GPT-12/2025-GPT-12.pdf`	present
2025-GPT-13	OpenAI ChatGPT (GPT-4o)	DEATH	A2	`/incidents/2025-GPT-13.pdf`	`/evidence/incidents/2025-GPT-13/2025-GPT-13.pdf`	present

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-MAI-01	Meta AI (Instagram/WhatsApp/Facebook)	UNSAFE_OUTPUT	A1	https://www.commonsensemedia.org/ai-ratings/meta-ai-risk-assessment; https://www.commonsensemedia.org/sites/default/files/featured-content/files/csm-ai-risk-assessment-metaai-08152025.pdf; `/incidents/2025-MAI-01.pdf`	`/evidence/incidents/2025-MAI-01/2025-MAI-01.pdf`	present
2025-MHB-01	Mental health chatbots (29 agents)	UNSAFE_OUTPUT	A1	Pichowicz, Kotas & Piotrowski (2025), Scientific Reports 15:31652, DOI: 10.1038/s41598-025-17242-4; https://doi.org/10.1038/s41598-025-17242-4; https://www.nature.com/articles/s41598-025-17242-4	https://doi.org/10.1038/s41598-025-17242-4	external_only

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-MLP-01	U.S. Senate Judiciary Committee, Subcommittee on Crime and Counterterrorism	HARM_EXPOSURE	A1	https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots; /incidents/2025-MLP-01_Raine_Testimony.pdf; /incidents/2025-MLP-01_Garcia_Testimony.pdf; /incidents/2025-MLP-01_Doe_Testimony.pdf; /incidents/2025-MLP-01_Torney_Testimony.pdf	/evidence/incidents/2025-MLP-01/2025-MLP-01_Raine_Testimony.pdf	present
2025-MLP-03	Multi-LLM red-team (6 models)	UNSAFE_OUTPUT	A1	https://arxiv.org/pdf/2507.02990.pdf; /incidents/2025-MLP-03.pdf	/evidence/incidents/2025-MLP-03/2025-MLP-03.pdf	present
2025-NOM-01	Nomi AI (Glimpse AI)	UNSAFE_OUTPUT	B	https://incidentdatabase.ai/cite/1041/; https://futurism.com/ai-girlfriend-encouraged-suicide	https://incidentdatabase.ai/cite/1041/	external_only

Table 15 continued

Incident ID	Platform	Outcome	Grade	Canonical primary citation / evidence anchor	Evidence route	Archive status
2025-REP-01	Replika (Luka Inc.)	HARM_EXPOSURE	A2	https://techjusticelaw.org/wp-content/uploads/2025/01/Complaint-and-Petition-for-Investigation-Re-Replika.pdf; `/incidents/2025-REP-01.pdf`	`/evidence/incidents/2025-REP-01/2025-REP-01.pdf`	present
2025-THR-01	Therapy chatbots (multi-app evaluation)	UNSAFE_OUTPUT	A1	Moore et al. (2025), FAccT '25, DOI: 10.1145/3715275.3732039; https://doi.org/10.1145/3715275.3732039; https://facctconference.org/static/docs/facct2025-206archivalpdfs/facct2025-final197-acmpaginated.pdf	https://doi.org/10.1145/3715275.3732039	external_only

F.2 Policy Document Source Index

The 16 policy rows are directly indexed in `policy_crosswalk_appendix.md`. For quick lookup, the public-document identifiers are:

Two Meta canonical URLs are retained even though `ai.meta.com` required login during the March 25, 2026 release audit; corresponding source-capture metadata are preserved in `policy_source_index.csv`.

Record ID	Document	Date	URL
ANTHROPIC-2025-01	How people use Claude for support, advice, and companionship	2025-06-27	https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship
ANTHROPIC-2025-02	Building safeguards for Claude	2025-08-12	https://www.anthropic.com/news/building-safeguards-for-claude
ANTHROPIC-2025-03	Protecting the well-being of our users	2025-12-18	https://www.anthropic.com/news/protecting-well-being-of-users
ANTHROPIC-2025-04	Sharing our compliance framework for California's Transparency in Frontier AI Act	2025-12-19	https://www.anthropic.com/news/compliance-framework-SB53
ANTHROPIC-2026-01	Anthropic's Responsible Scaling Policy: Version 3.0	2026-02-24	https://www.anthropic.com/news/responsible-scaling-policy-v3
GOOGLE-2024-01	Generative AI Prohibited Use Policy	2024-12-17	https://policies.google.com/terms/generative-ai/use-policy
GOOGLEDEEPMIND-2025-01	Frontier Safety Framework Report - Gemini 3 Pro (November, 2025) v2	2025-11	https://deepmind.google/models/fsf-reports/gemini-3-pro/
GOOGLEDEEPMIND-2025-02	Gemini 3 Pro - Model Card	2025-12	https://deepmind.google/models/model-cards/gemini-3-pro
META-2023-01	Llama 2 Acceptable Use Policy	2023-07-18	https://ai.meta.com/llama/use-policy/
META-2025-01	Code World Model Preparedness Report	2025-09-24	https://ai.meta.com/research/publications/code-world-model-preparedness-report/
OPENAI-2025-01	Preparedness Framework (Version 2)	2025-04-15	https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf
OPENAI-2025-02	Addendum to GPT-5 System Card: Sensitive Conversations	2025-10-27	https://cdn.openai.com/pdf/3da476af-b937-47fb-9931-88a851620101/addendum-to-gpt-5-system-card-sensitive-conversations.pdf
OPENAI-2025-03	Model Spec (2025/12/18)	2025-12-18	https://model-spec.openai.com/2025-12-18.html

Record ID	Document	Date	URL
OPENAI-2026-01	GPT-5.3 Instant System Card	2026-03-03	https://deploymentsafety.openai.com/gpt-5-3-instant/gpt-5-3-instant.pdf
XAI-2025-01	Grok 4.1 Model Card	2025-11-17	https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf
XAI-2025-02	xAI Frontier Artificial Intelligence Framework	2025-12-30	https://data.x.ai/2025-12-31-xai-frontier-artificial-intelligence-framework.pdf

Data Availability and Declarations

Data Availability

Version 2.0 study materials are available at https://selfharm.ai/downloads/. The downloads endpoint is intended to resolve at public release; pre-release review copies may encounter an inactive URL before DOI minting and site publication. They include the coded incident registry, the analysis companion, the coded policy registry, schemas, coding manuals, search-execution log, screening and source-capture ledgers, incident and policy audit scaffolds, archived evidence files where redistribution was feasible, and standalone copies of the appendices. Citation metadata, provenance notes, and integrity documentation are also available there.

Some records are intentionally routed to external primary sources rather than archived locally, and some legal materials remain paywalled or externally hosted. No private platform logs or non-public human-subject materials are included.

Ethics Review

This study used only publicly available records, court filings, hearing materials, public journalism, benchmark reports, and provider-issued governance documents. No direct participant recruitment, private records, or non-public human-subject data were used. On that basis, the work did not require IRB or human-subjects review under the applicable public-records standard.

Funding

This work received no external funding.

Competing Interests

The author declares no competing interests.

Sensitive-Material Handling

This article discusses suicide and self-harm but does not reproduce operational instructions. Means-specific details are redacted or omitted, quotations are minimized, and the focus remains on system behavior, evidentiary discipline, and prevention-oriented evaluation design. Because the corpus includes minors and deaths, the article relies only on already public materials and does not introduce new identifying detail beyond what is already part of the source record.

AI-Use Statement

Where generative AI assistance was used, it was not used to identify incidents, select sources, assign incident taxonomy codes, determine reliability grades, extract evidentiary claims, or resolve interpretive disputes. No AI-generated text was treated as evidence. AI assistance was used only for drafting, rephrasing, table formatting, and revision support, with all retained text checked against the cited source material and coded materials.

References

Artificial Intelligence Incident Database. (n.d.-a). *CSETv1 charts*. Retrieved March 2026, from https://incidentdatabase.ai/taxonomies/csetv1/

Artificial Intelligence Incident Database. (n.d.-b). *Editor's guide*. Retrieved March 2026, from https://incidentdatabase.ai/editors-guide/

Artificial Intelligence Incident Database. (n.d.-c). *GMF charts*. Retrieved March 2026, from https://incidentdatabase.ai/taxonomies/gmf/

Anthropic. (2025, June 27). *How people use Claude for support, advice, and companionship*. https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship

Anthropic. (2025, August 12). *Building safeguards for Claude*. https://www.anthropic.com/news/building-safeguards-for-claude

Anthropic. (2025, December 18). *Protecting the well-being of our users*. https://www.anthropic.com/news/protecting-well-being-of-users

Anthropic. (2025, December 19). *Sharing our compliance framework for California's Transparency in Frontier AI Act*. https://www.anthropic.com/news/compliance-framework-SB53

Anthropic. (2026, February 24). *Anthropic's Responsible Scaling Policy: Version 3.0*. https://www.anthropic.com/news/responsible-scaling-policy-v3

Centers for Disease Control and Prevention. (2024, November 29). *Youth mental health: The numbers*. https://www.cdc.gov/healthy-youth/mental-health/mental-health-numbers.html

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C. W., Shan, C. Y., & Wadman, K. (2025). *How people use ChatGPT* (Working Paper No. 34255). National Bureau of Economic Research. https://www.nber.org/papers/w34255

Faverio, M., & Sidoti, O. (2025, December 9). *Teens, social media and AI chatbots 2025*. Pew Research Center. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/

Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. May 21, 2025) (order on motions to dismiss).

Gavalas v. Google LLC and Alphabet Inc., No. 5:26-cv-01849 (N.D. Cal. filed March 4, 2026).

Google. (2024, December 17). *Generative AI Prohibited Use Policy*. https://policies.google.com/terms/generative-ai/use-policy

Google DeepMind. (2025, November). *Frontier Safety Framework Report - Gemini 3 Pro (November, 2025) v2*. https://deepmind.google/models/fsf-reports/gemini-3-pro/

Google DeepMind. (2025, December). *Gemini 3 Pro - Model Card*. https://deepmind.google/models/model-cards/gemini-3-pro

Meta. (2023, July 18). *Llama 2 Acceptable Use Policy*. https://ai.meta.com/llama/use-policy/

Meta. (2025, September 24). *Code World Model Preparedness Report*. https://ai.meta.com/research/publications/code-world-model-preparedness-report/

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). *Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers*. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 599–627). https://doi.org/10.1145/3715275.3732039

OECD. (2025). *Towards a common reporting framework for AI incidents* (OECD Artificial Intelligence Papers No. 34). OECD Publishing. https://doi.org/10.1787/f326d4ac-en

OECD.AI. (n.d.-a). *Overview and methodology of the AI Incidents and Hazards Monitor*. Retrieved March 2026, from https://oecd.ai/en/incidents-methodology

OECD.AI. (n.d.-b). *Name it to tame it: Defining AI incidents and hazards*. Retrieved March 2026, from https://oecd.ai/en/wonk/defining-ai-incidents-and-hazards

Ofcom. (2025, August 22). *Protecting people from online suicide and self-harm material*. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/protecting-people-from-online-suicide-and-self-harm-material

OpenAI. (2025, April 15). *Preparedness Framework (Version 2)*. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

OpenAI. (2025, October 27). *Addendum to GPT-5 System Card: Sensitive Conversations*. https://cdn.openai.com/pdf/3da476af-b937-47fb-9931-88a851620101/addendum-to-gpt-5-system-card-sensitive-conversations.pdf

OpenAI. (2025, December 18). *Model Spec (2025/12/18)*. https://model-spec.openai.com/2025-12-18.html

OpenAI. (2026, March 3). *GPT-5.3 Instant System Card*. https://deploymentsafety.openai.com/gpt-5-3-instant/gpt-5-3-instant.pdf

Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). *Performance of mental health chatbot agents in detecting and managing suicidal ideation*. *Scientific Reports*, 15, 31652. https://doi.org/10.1038/s41598-025-17242-4

Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

Spittal, M. J., et al. (2025). *Can suicide and self-harm in children and adolescents be predicted by mental health and social care contact? A systematic review and meta-analysis*. *PLOS Medicine*, 22(8), e1004581. https://doi.org/10.1371/journal.pmed.1004581

Tricco, A. C., Lillie, E., Zarin, W., et al. (2018). *PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation*. *Annals of Internal Medicine*, 169(7), 467-473. https://doi.org/10.7326/M18-0850

U.S. Senate Judiciary Committee, Subcommittee on Crime and Counterterrorism. (2025, September 16). *Examining the harm of AI chatbots*. https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots

World Health Organization. (2019, June 24). *Preventing suicide: A resource series*. https://www.who.int/publications/i/item/preventing-suicide-a-resource-series

xAI. (2025, November 17). *Grok 4.1 Model Card*. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf

xAI. (2025, December 30). *xAI Frontier Artificial Intelligence Framework*. https://data.x.ai/2025-12-31-xai-frontier-artificial-intelligence-framework.pdf