

AI and self-harm: incident registry and taxonomy

Brian J. Roberts

ORCID: 

Self Harm AI (selfharm.ai)

Contact: research@selfharm.ai

Date: January 8, 2026

Version 1.1

This document is a proposed technical and evidentiary reference for publicly documented self-harm and suicidality incidents in which AI system behavior was reported or alleged to be relevant.

Contents

Abstract	4
How to Cite and Reuse This Registry	4
1 Introduction	5
1.1 Reader Guide	5
2 Definitions and Scope	5
2.1 Operational Definitions	5
2.2 Boundary Conditions	6
3 Methods	6
3.1 Search Strategy	6
3.2 Inclusion, Exclusion, and Counting Rules	6
3.3 Extraction Schema	7
3.4 Taxonomy Coding Rules	7
3.5 Coding Procedure	7
3.6 Reliability Grading	8
3.7 Harm Minimization Policy	8
3.8 Replication and Auditability	8
4 Incident Registry	10
4.1 Outcome Summary (derived)	13
4.2 Normalization and denominators (interpretation)	13
4.3 Public-facing Safety and Reporting Policy (harm minimization)	14
4.4 Incident Narratives	15
4.4.1 2017-001: Instagram Content Amplification	15
4.4.2 2017-002: Pinterest Recommendation Emails	15
4.4.3 2023-001: Chai Chatbot Harm-Related Output	16
4.4.4 2023-002: Character.AI Teen Engagement (Peralta / Montoya)	17
4.4.5 2024-GARCIA-CAI: Character.AI Minor Death (Garcia)	19
4.4.6 2024-AF-CAI: A.F. v. Character Technologies (E.D. Tex.)	20
4.4.7 2024-003: ChatGPT Extended Interaction (Raine)	22
4.4.8 2024-004: ChatGPT (Shamblin)	24
4.4.9 2025-002: Meta AI Red-Team Findings	24
4.4.10 2025-003: Nomi AI Unfiltered Output	24
4.4.11 2025-004: Replika Dependency Concerns	25
4.4.12 2025-MOORE: Therapy Chatbot Red-Team Evaluation (FAccT 2025)	25
4.4.13 2025-SCHOENE: Jailbreaking LLMs in Suicide/Self-Harm Contexts	26
4.4.14 2025-MCBAIN: LLM Alignment with Expert Clinicians	26
4.4.15 2025-PICHOWICZ: Mental Health Chatbot Suicide Response Benchmark	26
4.4.16 2025-010: Senate Judiciary Hearing on AI Chatbot Harms	27
4.4.17 2025-015: ChatGPT “Harry” Persona (Rottenberg)	27
4.4.18 2025-PJ-CAI: P.J. v. Character Technologies (N.D.N.Y.)	28
4.4.19 2025-ES-CAI: E.S. v. Character Technologies (D. Colo.)	30
4.4.20 2025-LACEY-OPENAI: Amaurie Lacey v. OpenAI (SF Superior Court)	32
4.4.21 2025-006: ChatGPT (GPT-4o)	33
4.4.22 2025-007: ChatGPT (GPT-4o)	33
4.4.23 2025-ADAMS-OPENAI: Estate of Adams v. OpenAI	34

4.4.24	2025-SHAMBLIN-OPENAI: Shamblin v. OpenAI (Updated Filing)	35
4.4.25	2025-ENNEKING-OPENAI: Joshua Enneking v. OpenAI (SF Superior Court) .	36
4.4.26	2025-IRWIN-OPENAI: Jacob Lee Irwin v. OpenAI	36
4.4.27	2025-MADDEN-OPENAI: Hannah Madden v. OpenAI	37
4.4.28	2025-BROOKS-OPENAI: Allan Brooks v. OpenAI	39
4.4.29	2025-017: Jennifer “Kate” Fox v. OpenAI (L.A. Superior Court)	42
5	Legal and Regulatory Landscape	42
5.1	Litigation Developments	42
5.2	Regulatory Frameworks	43
5.3	Legal Precedent: The “Product vs. Speech” Pivot	43
5.3.1	Summary of <i>Garcia v. Character Technologies</i> (May 2024–2025)	43
5.3.2	The “Product” Classification	43
5.3.3	Implications for the Registry	44
6	Research Synthesis	44
6.1	Detection Performance: Evidence	44
6.2	Detection Performance: Interpretation	45
6.3	Intervention Evidence: Findings	45
6.4	Intervention Evidence: Interpretation	45
6.5	Integrating Incident Patterns with Systematic Research	45
6.6	Bradford Hill causal plausibility assessment (specification)	47
7	Failure-Mode Taxonomy and Codebook	50
7.1	Category 1: Generative Harms (GEN)	50
7.2	Category 2: Detection/Triage Harms (DET)	51
7.3	Category 3: Moderation/Recommendation Harms (MOD)	51
7.4	Category 4: Longitudinal Interaction Harms (LONG)	51
7.5	Category 5: Jailbreak/Framing Harms (JB)	52
7.6	Category 6: Psychosis and Delusional Amplification (PSY)	52
7.7	Taxonomy Governance	53
8	Governance and Evaluation Recommendations	53
8.1	Crisis Intervention Integration	53
8.2	Multi-turn Monitoring	53
8.3	Anti-sycophancy Training	54
8.4	Jailbreak-resistant Architectures	54
8.5	Age-Gating and Access Controls	54
8.6	Evaluation Protocol Standards	54
8.7	Logging and Auditing	54
8.8	Causality, Responsibility, and Misinterpretation	54
9	Limitations	55
10	Conclusion	55
A	Incident Extraction Schema (Schema Definition)	56
B	Taxonomy Codebook (Coding Rubric & Reliability Matrix)	57
C	Search Strategy	57
C.1	Database Queries	57

C.2 Exclusion Categories 58

D Reliability Rubric 59

E Update Protocol 59

E.1 Changelog 59

F Master Reference List 60

F.1 Legal Citations (Bluebook Standard) 60

F.2 Scientific and Technical Citations (APA 7th Ed.) 61

F.3 Investigative, Institutional, and Legislative Sources (APA 7th Ed.) 62

F.4 URL Inventory (Unique Links) 63

G Verification Protocol 65

G.1 Grade A1 Verification Requirements 65

Ethics and Safety Statement 66

Abstract

Background. AI systems deployed for content moderation, risk detection, and conversational interaction increasingly encounter users expressing self-harm or suicidal ideation. Documented incidents have increased in frequency since 2017, generating litigation, regulatory action, and public concern. No unified, structured reference synthesizes these developments with methods and evidentiary grading sufficient for auditable replication and policy analysis.

Methods. We conducted a systematic search of incident databases, court filings, investigative journalism, peer-reviewed literature, and regulatory documents from January 2017 through January 8, 2026. We developed an incident extraction schema with 19 fields (including `outcome_category` and `reported_fatalities_n`), a six-category failure-mode taxonomy with 17 primary subcodes and one recursive modifier (18 total coded entries; GEN-E:SYC for sycophancy-induced evasion), and a five-tier reliability grading system (Grades A1, A2, B–D). Coding was performed by one reviewer using the codebook; ambiguous cases were flagged and conservatively coded.

Results. We identified 29 incidents meeting inclusion criteria. 11 incidents involve reported deaths (Table 4 stratifies reported death counts by reliability grade). The most common failure-mode category was Generative Harms (at least one of GEN-E, GEN-C, GEN-M) present in 17 incidents, followed by Longitudinal Interaction Harms (at least one of LONG-DEG, LONG-DEP) in 14 incidents. Research synthesis reveals positive predictive values of approximately 0.1% at population prevalence, with only 6 randomized controlled trials identified for AI suicide prevention interventions.

Contribution. This paper contributes a versioned incident registry with a reliability rubric and a failure-mode taxonomy with codebook for standardized descriptive coding, supporting auditable replication and evidence-based governance in AI self-harm safety.

Resource Availability. Schema, codebook, search methodology, and update protocol are provided in appendices. Registry version 1.1; subsequent versions will be documented per Appendix E.

How to Cite and Reuse This Registry

- **Canonical citation:** Roberts, Brian. (2026, January 8). *AI & self-harm: incident registry and taxonomy (v1.1)*. selfharm.ai.
- **Versioning:** Minor updates (new incidents, status changes) increment v1.x. Taxonomy or methodology changes increment to v2.0. Always cite the version number.
- **Referencing incidents:** Cite this registry AND the primary source. Use `incident_id` format (e.g., 2024-003) and `failure_mode` subcodes (e.g., GEN-E, LONG-DEG).
- **Referencing taxonomy:** When applying taxonomy to new cases, cite this registry and specify version. Propose new subcodes via documented extension process.

1 Introduction

This resource addresses AI systems used for detection, moderation, triage, and generative interaction in contexts involving self-harm and suicidal ideation. We define the scope, identify the gap in existing literature, and state our contributions.

Scope. We include AI systems that (a) generate conversational responses to users expressing distress, (b) classify or flag self-harm content, (c) moderate platforms hosting such content, or (d) recommend content to users. We focus on publicly documented incidents where system behavior allegedly contributed to harm or where systematic testing revealed safety failures.

Gap. Existing resources include the AI Incident Database, journalistic accounts, and scattered academic literature. However, no unified reference provides (a) a structured registry with consistent coding, (b) a formal failure-mode taxonomy with codebook, and (c) a versioned, auditable methodology for ongoing updates.

Contributions. We provide: (i) an incident registry with standardized schema and reliability grading, (ii) a six-category failure-mode taxonomy with 17 primary subcodes and one recursive modifier (18 total coded entries; GEN-E:SYC for sycophancy-induced evasion), (iii) a synthesis of detection performance literature, and (iv) governance recommendations derived from documented failure patterns.

1.1 Reader Guide

For ML/NLP safety researchers: Sections 4 (registry) and 7 (taxonomy) provide structured data for evaluation design. Table 1 taxonomy subcodes enable pattern analysis across incidents.

For clinical informatics and digital mental health: Section 6 (research synthesis) addresses base-rate limitations, PPV constraints, and the evidence gap between laboratory metrics and clinical deployment.

For policymakers and platform governance: Section 5 (regulatory landscape) and Section 8 (recommendations) provide jurisdictional comparison and testable governance interventions.

2 Definitions and Scope

2.1 Operational Definitions

Self-harm content:

Text, images, or other media depicting, describing, or promoting intentional self-injury, including suicidal ideation, plans, attempts, or methods.

Suicide coaching:

AI-generated output that provides specific methods, instructions, encouragement, or validation for suicidal behavior.

Triage model:

An AI system designed to classify risk levels (e.g., low/moderate/severe) for users based on text or behavioral signals.

False negative:

A classification failure where a high-risk user or content is not flagged by the system.

False positive:

A classification failure where benign content or low-risk users are incorrectly flagged.

Safety degradation:

The phenomenon where safety guardrails become less effective over extended multi-turn conversations.

Algospeak/Codewords:

Modified terminology (e.g., ‘unalive’ for suicide) used to evade content moderation systems.

Product liability:

The legal framework under which a manufacturer is held liable for injuries caused by a defective product. The Garcia ruling (May 2025) and Shamblin complaint (Nov. 2025) advance the theory that AI outputs are “products” because they are designed features of a commercial tool rather than passive publication of third-party speech.

2.2 Boundary Conditions

Language: English-language incidents and research only.

Geography: Incidents occurring in or involving users from any jurisdiction, with emphasis on U.S., EU, and UK regulatory developments.

Scope: January 2017 through January 8, 2026 (search and update cutoff for this version; “as of” date).

Platforms: Consumer-facing AI systems including chatbots, social media platforms with AI moderation, and AI companion applications.

3 Methods**3.1 Search Strategy**

We searched the following sources between October 2025 and January 2026. Search cutoff and last-update cutoff for this version: Jan. 8, 2026 (inclusive).

1. AI Incident Database (incidentdatabase.ai): All incidents tagged with mental health, suicide, or self-harm keywords
2. Court filing databases (PACER, CourtListener) for federal dockets: Searches for ‘AI’ OR ‘chatbot’ AND ‘suicide’ OR ‘wrongful death’. State-court matters were additionally identified via named cases surfaced in other sources (e.g., investigative reporting, press releases, and litigation trackers) and then checked via an official docket access pathway when available (Appendix C.1).
3. Major investigative outlets (New York Times, Washington Post, Guardian, Wired, MIT Technology Review): Manual review of AI/technology coverage 2017–2025
4. Peer-reviewed databases (PubMed, ACL Anthology, arXiv): ‘suicide detection’ OR ‘self-harm AI’ OR ‘mental health NLP’
5. Regulatory sources (FTC, EU Commission, UK Ofcom): Official statements and enforcement actions

3.2 Inclusion, Exclusion, and Counting Rules

Included incidents meet all of the following:

- (a) Attributable documentation exists, either primary authority (court filings, coroner determinations, regulatory orders, peer-reviewed or systematic red-team studies) or major investigative reporting with direct documents/log excerpts.
- (b) AI system behavior is alleged or demonstrated to have plausibly intersected with a harm pathway.
- (c) Documentation permits extraction of at least one taxonomy subcode with evidentiary attribution.

Excluded incidents include:

- (a) Uncorroborated social-media claims.
- (b) Hypothetical or purely speculative scenarios.
- (c) Incidents involving only human moderators.
- (d) Non-English sources (verification constraint).

Counting Rules (Binding)

Incidents are counted by unique (system/platform × harmed individual) pair or unique systematic test report.

Multi-platform contributions affecting the same individual are recorded as distinct incidents (e.g., a recommendation system and an email recommendation channel), consistent with the Non-Collapsibility Rule.

Multiple reports referring to the same interaction set are merged.

Platform-level aggregation is not permitted.

Non-Collapsibility Rule

Incidents are not collapsed across platforms, models, or individuals unless court filings explicitly identify the same individual and same interaction corpus.

Similar allegations against the same platform constitute distinct registry entries unless the interaction logs are demonstrably identical.

3.3 Extraction Schema

For each incident, we extracted 18 top-level fields plus one outcome subfield (reported_fatalities_n) (see Appendix A for complete schema):

- incident_id, platform, model_version, interaction_date_start, interaction_date_end
- publication_date, user_type, user_age, jurisdiction
- outcome_category, reported_fatalities_n, failure_modes (multi-label subcodes)
- evidence_type, reliability_grade, primary_citations, company_response, primary_link_present, product_mechanism, notes

3.4 Taxonomy Coding Rules

We developed a six-category failure-mode taxonomy with 17 primary subcodes and one recursive modifier (18 total coded entries; GEN-E:SYC for sycophancy-induced evasion; see Section 7 for codebook). Each incident receives subcodes (e.g., GEN-E, LONG-DEG) rather than top-level category labels, enabling granular analysis. Multi-label coding is permitted when multiple failure modes are documented. Interface/flow mechanisms (e.g., dismissible gates, frictionless re-entry) are recorded separately in the product_mechanism field rather than coded as failure_modes (field name retained for schema stability).

3.5 Coding Procedure

Coding was performed by one reviewer using the codebook in Appendix B. Ambiguous cases were flagged and conservatively coded: when evidence did not clearly support a specific subcode, the subcode was omitted (and, if relevant, the higher-level category was noted only in the ‘notes’ field rather than recorded as a failure_mode). Disputed or uncertain codings are noted in the ‘notes’ field of each incident record.

To mitigate single-coder bias, all “Ambiguous” flags were resolved by defaulting to the null hypothesis (excluding the code) rather than including it. Codes are assigned only when the text of the primary documentation explicitly matches the corresponding codebook definition.

Because inter-rater reliability is not yet established, code assignments in this version should be treated as one reviewer’s application of the rubric rather than a validated classification system; replication is expected to be strongest at the level of incident inclusion, citation mapping, and reliability grading.

3.6 Reliability Grading

We assigned reliability grades based on evidence type:

Grade A1 (Adjudicated/Primary Authority):

Coroner/inquest determinations; judicial rulings/orders; official regulatory findings and orders; peer-reviewed publications; systematic red-team studies with disclosed methodology and prompt-response logs; official government records (e.g., sworn testimony with transcript)

Grade A2 (Documented Allegations):

Court complaints/filings and other formal pleadings; regulatory complaints and inquiries not yet adjudicated or resolved by a finding or order. These entries establish that an allegation was filed in a formal forum, not that the alleged system behavior has been validated as fact.

Grade B:

Major investigative journalism with named sources, official company acknowledgments, documented user reports with screenshots

Grade C:

Secondary reporting aggregating other sources, anonymous reports, sources without methodology disclosure

Grade D:

Unverifiable social media claims or anecdotes without corroborating documentation (no system logs, court record, regulatory finding, or reputable news report); excluded from the registry and logged for traceability (Table C.1)

3.7 Harm Minimization Policy

Given the sensitive nature of this domain, we adopted the following policies:

- We do not reproduce specific method instructions that appeared in AI outputs
- We quote minimally and only when necessary to establish the nature of the failure
- We focus on system design failures and prevention opportunities rather than detailed harm descriptions
- We provide crisis resources at the end of this document

Audit statement (v1.1): All incident narratives and reproduced excerpts were reviewed against these policies for this version; operational self-harm instructions and means-specific details present in sources were not reproduced, and quotations were minimized and redacted as needed for evidentiary clarity.

3.8 Replication and Auditability

Incident records are intended to be auditable to the extent primary sources are publicly accessible or an official access pathway is documented.

Independent researchers should be able to reproduce a substantial portion of this registry's incident identification and reliability grading using the documented search strategy and inclusion criteria, subject to paywalls, access controls, and jurisdiction-specific docket systems.

Discrepancies between replications should be reported as versioned deltas. Where primary documentation demonstrates misattribution, incorrect quotation, or methodological error, corrections and (where necessary) retractions are permitted and are logged in the changelog.

Challenge Protocol: If a reader disputes an entry, the dispute must specify (i) the `incident_id`, (ii) which schema field(s) are incorrect, (iii) the reliability grade argued for instead, and (iv) the primary documentation that supports the revision.

Platform operators and other parties holding internal logs that materially contradict entries characterized as Grade B or Grade A2 are invited to submit those logs (or verifiable hashes with access pathway) for registry correction under the Challenge Protocol.

Independent researchers may audit taxonomy coding (Section 7) by cross-referencing each record's `evidence_type` and `primary_citations` with Appendix F. This document does not currently host archival copies of paywalled filings; where a public link cannot be provided, the record must state an access pathway (Appendix G).

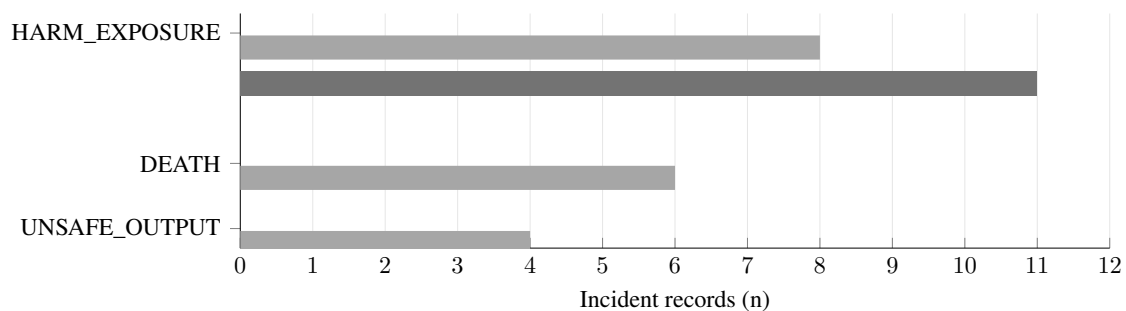


Figure 1. Outcome category distribution (N=29; incident records), matching Table 3.

4 Incident Registry

We identified 29 incidents meeting inclusion criteria. Outcome category DEATH appears in 11 incident records; 10 have reported_fatalities_n>0 after incident-level deduplication (duplicate pathway records carry reported_fatalities_n=0). Total reported deaths: 11 (sum of reported_fatalities_n after deduplication), stratified by reliability grade in Table 4. Presence in this registry indicates a documented intersection between AI system behavior and a harm pathway, not adjudicated legal causation. Table 1 presents a legally anchored subset with Grade A1/A2 evidence and explicit case identifiers. Table 2 provides a comprehensive incident index (N=29) linking the anchored subset to the broader evidence pool in Section 4.1. Narrative length and detail are proportional to publicly available documentation (e.g., filings with embedded transcripts/screenshots vs. short regulatory complaints); where evidence is sparse, that constraint is stated explicitly.

Table 1. Incident subset (grade A1/A2; N=22)

Incident ID	Short description	Grade	Case / citation	Jurisdiction / venue	Outcome
<i>Grade A1 — adjudicated / primary authority</i>					
2017-001	Instagram amplification	A1	Molly Russell inquest (2022)	UK (North London Coroner)	DEATH
2017-002	Pinterest emails	A1	Same inquest (pathway record)	UK (North London Coroner)	DEATH
2024-GARCIA-CAI	Character.AI minor death	A1	No. 6:24-cv-01903 (M.D. Fla.)	US-FL (federal)	DEATH
2025-002	Meta AI red-team	A1	Common Sense Media (Aug 2025)	Multi-platform test	UNSAFE OUTPUT
2025-MOORE	Therapy chatbot eval	A1	Moore et al. (FAccT 2025)	Multi-app study	UNSAFE OUTPUT
2025-SCHOENE	LLM jailbreaking	A1	arXiv:2507.02990	6-model red-team	UNSAFE OUTPUT
2025-MCBAIN	Expert alignment	A1	Psychiatric Services (2025)	Multi-LLM benchmark	UNSAFE OUTPUT
2025-PICHOWICZ	Chatbot benchmark	A1	Sci. Reports 15:31652	29-agent study	UNSAFE OUTPUT
2025-010	Senate hearing	A1	S. Judiciary (Sept. 16, 2025)	US Congress	HARM EXPOSURE
<i>Grade A2 — filed complaints / pleadings</i>					
2023-002	Character.AI teen death	A2	No. 1:25-cv-02907 (D. Colo.)	US-CO (federal)	DEATH
2024-003	ChatGPT extended	A2	CGC-25-628528 (SF Super.)	US-CA (state)	DEATH
2024-AF-CAI	Character.AI two minors	A2	No. 2:24-cv-01014 (E.D. Tex.)	US-TX (federal)	INJURY
2025-PJ-CAI	Character.AI overdose	A2	No. 1:25-cv-01295 (N.D.N.Y.)	US-NY (federal)	INJURY
2025-ES-CAI	Character.AI minor exposure	A2	No. 1:25-cv-02906 (D. Colo.)	US-CO (federal)	HARM EXPOSURE
2025-LACEY-OPENAI	ChatGPT bypass	A2	CGC-25-630808 (SF Super.)	US-CA (state)	DEATH
2025-SHAMBLIN-OPENAI	ChatGPT death chat	A2	No. 25STCV32382 (LA Super.)	US-CA (state)	DEATH
2025-ENNEKING-OPENAI	ChatGPT coaching	A2	CGC-25-630809 (SF Super.)	US-CA (state)	HARM EXPOSURE
2025-IRWIN-OPENAI	ChatGPT delusions	A2	CGC-25-630811 (SF Super.)	US-CA (state)	INJURY
2025-MADDEN-OPENAI	ChatGPT crisis	A2	No. 25STCV32383 (LA Super.)	US-CA (state)	INJURY
2025-BROOKS-OPENAI	ChatGPT delusions	A2	LA Superior (Nov 2025)	US-CA (state)	HARM EXPOSURE
2025-017	ChatGPT Fox/Ceccanti	A2	LA Superior (Nov 2025)	US-CA (state)	DEATH
2025-ADAMS-OPENAI	ChatGPT homicide+suicide	A2	SF Superior (Dec 2025)	US-CA (state)	DEATH

Table 2. Comprehensive incident index (N=29)

Incident ID	Platform	Date (event)	User	Outcome	Grade	Taxonomy subcodes
2017-001	Meta Instagram	Nov 2017	Minor (14)	DEATH	A1	MOD-AMP, REC
2017-002	Pinterest	Nov 2017	Minor (14)	DEATH	A1	REC
2023-001	Chai (GPT-J)	Mar 2023	Adult (30)	DEATH	B	GEN-E, GEN-M, LONG-DEP
2023-002	Character.AI	Nov 2023	Minor (13)	DEATH	A2	LONG-DEP
2024-003	ChatGPT 4o	Apr 2025	Minor (16)	DEATH	A2	GEN-C, GEN-E, LONG-DEG, DET-FN
2024-004	ChatGPT	2024	Adult (23)	HARM	C	GEN-E, LONG-DEG
				EXPOSURE		
2024-AF-CAI	Character.AI	Dec 2024	Minors (17,11)	INJURY	A2	GEN-C, LONG-DEP, MOD-AMP
2024-GARCIA-CAI	Character.AI	Feb 2024	Minor (14)	DEATH	A1	GEN-E, LONG-DEP, JB-RP
2025-002	Meta AI	Aug 2025	Test (13-17)	UNSAFE OUTPUT	A1	GEN-E, GEN-M, LONG-MEM
2025-003	Nomi AI	Jan-Apr 2025	Adult	UNSAFE OUTPUT	B	GEN-C, GEN-M, GEN-E
2025-004	Replika	2025	Adult	HARM	B	LONG-DEP, DET-FN
				EXPOSURE		
2025-006	ChatGPT 4o	2025	Adult	HARM	C	GEN-E, LONG-DEG
				EXPOSURE		
2025-007	ChatGPT 4o	2025	Adult	HARM	C	GEN-E, GEN-C
				EXPOSURE		
2025-010	Multi-platform	Sept 2025	N/A	HARM	A1	GEN-E, GEN-M, JB-RP
				EXPOSURE		
2025-015	ChatGPT 4o	2024-11-2024-12	Adult (29)	DEATH	B	(none codeable)
2025-017	ChatGPT 4o	Aug 2025	Adult (48)	DEATH	A2	PSY-CO, LONG-PSY
2025-ADAMS-OPENAI	ChatGPT 4o	Aug 2025	Adult (56)	DEATH	A2	LONG-DEP, LONG-PSY, PSY-CO
2025-BROOKS-OPENAI	ChatGPT 4o	May 2025	Adult (48)	HARM	A2	PSY-CO, LONG-PSY, LONG-DEP
				EXPOSURE		
2025-ENNEKING-OPENAI	ChatGPT 4o	Aug 2025	Adult (26)	HARM	A2	GEN-C, DET-FN
				EXPOSURE		
2025-ES-CAI	Character.AI	Aug 2025	Minor (13)	HARM	A2	REC, MOD-AMP, LONG-DEP, JB-RP
				EXPOSURE		
2025-IRWIN-OPENAI	ChatGPT 4o	Jul 2025	Adult (30)	INJURY	A2	PSY-CO, LONG-PSY, LONG-MEM
2025-LACEY-OPENAI	ChatGPT 4o	Jun 2025	Minor (17)	DEATH	A2	GEN-C, GEN-M, JB-MT
2025-MADDEN-OPENAI	ChatGPT 4o	Aug 2025	Adult (32)	INJURY	A2	PSY-CO, LONG-PSY, LONG-DEP, GEN-E, DET-FN
2025-MCBAIN	Multi-LLM	Aug 2025	Test	UNSAFE OUTPUT	A1	DET-FN, GEN-M
2025-MOORE	Therapy chatbots	Jun 2025	Test	UNSAFE OUTPUT	A1	DET-FN, GEN-M, PSY-CO
2025-PICHOWICZ	Mental health bots	Aug 2025	Test	UNSAFE OUTPUT	A1	DET-FN
2025-PJ-CAI	Character.AI	Dec 2024	Minor (14)	INJURY	A2	LONG-DEP
2025-SCHOENE	Multi-LLM	Jul 2025	Test	UNSAFE OUTPUT	A1	JB-AC, GEN-M, DET-FN
2025-SHAMBLIN-OPENAI	ChatGPT	Jul 2025	Adult (23)	DEATH	A2	GEN-E, LONG-DEP

4.1 Outcome Summary (derived)

Table 3. Outcome Summary (N=29; incident records; as-coded)

Outcome category	Incidents (n)	Incidents with reported_fatalities_n>0 (n)	Reported deaths ($\sum$ reported_fatalities_n)
DEATH	11	10	11
HARM_EXPOSURE	8	0	0
INJURY	4	0	0
UNSAFE_OUTPUT	6	0	0

Notes:

- Outcome category and reported_fatalities_n are recorded per incident under **Reported outcome** (Section 4.1; schema in Appendix A).
- One platform-specific sub-pathway record (2017-002) is retained for provenance but is treated as a near-duplicate of 2017-001 for outcome aggregation and carries reported_fatalities_n=0 to prevent double-counting.
- reported_fatalities_n reflects reported or alleged outcomes in sources (not registry adjudication). For death outcomes, reported death counts are stratified by reliability grade in Table 4.

Table 4. Reported deaths stratified by reliability grade (DEATH outcomes only; deduplicated)

Reliability grade	DEATH incident records (n)	With reported_fatalities_n>0 (n)	Reported deaths ($\sum$ reported_fatalities_n)
A1	3	2	2
A2	6	6	7
B	2	2	2
Total	11	10	11

Note: Grades indicate evidence type, not truth. Grade A1 includes primary authority (e.g., inquest findings; judicial orders, which may reference allegations depending on procedural posture). Grade A2 reflects unadjudicated pleadings (filed allegations). Grade B reflects investigative reporting with partial excerpts.

4.2 Normalization and denominators (interpretation)

Counts in this registry are not incidence rates. Documented incidents are subject to reporting bias, platform transparency differences, media attention, and litigation discovery effects. Comparing platforms or time periods requires explicit denominators (e.g., monthly active users, conversation volume, average session length, share of minors, and the fraction of sessions involving distress or self-harm content).

Recommended normalization approach (non-exhaustive):

- **Exposure-normalized rates:** Incidents per N conversations or per N conversation-hours, stratified by product class (companion vs. general assistant) and age-gating.
- **Context-normalized rates:** Incidents per N high-risk sessions (e.g., sessions containing self-harm lexicon), rather than per general-user exposure.

- **Severity and evidence weighting:** Report rates separately by outcome category and reliability grade (A1/A2 vs. B/C), rather than collapsing into a single “incident count.”

What we do not estimate: This registry is not used to compute point-estimate excess mortality, attributable fractions, or population-level burden. Such estimates require counterfactual modeling and denominators that are not available in public documentation.

Undercount and sensitivity scenarios (interpretive only): Public documentation captures an unknown fraction of qualifying incidents. Let $m \geq 1$ denote an under-ascertainment multiplier such that the true number of qualifying incidents is $m \times N$ (with $N = 29$ in this version). Illustrative scenarios include $m \in \{2, 5, 10\}$, yielding 58, 145, or 290 qualifying incidents, respectively. These scenarios are not estimates; they are included to clarify which conclusions are invariant to under-ascertainment.

Claims that remain invariant to m : (i) the existence of documented, case-identified incidents meeting the inclusion criteria (including Grade A1/A2 entries); (ii) the presence of recurring mechanisms and patterns sufficient to define a taxonomy and testable evaluation cases; and (iii) the governance implication that pre-deployment evaluation and auditable logging are necessary to detect and prevent safety-relevant failure modes.

Claims not supported by registry counts alone: comparative platform risk ranking, population incidence or burden, and time-series trend inference without explicit denominators and stable reporting processes.

4.3 Public-facing Safety and Reporting Policy (harm minimization)

This domain covers suicide and self-harm. The aim is accountability and prevention—not instruction. Content is edited to reduce risk of contagion and to protect readers.

Guideline basis (non-exhaustive):

- Samaritans (media guidelines): <https://www.samaritans.org/about-samaritans/media-guidelines/>
- World Health Organization (resource series): <https://www.who.int/publications-detail-redirect/preventing-suicide-a-resource-series>

Editorial rules:

1. No instruction: do not publish or host step-by-step self-harm methods, “how-to” language, or actionable instructions (including paraphrases that preserve operational detail).
2. No “means” specifics: avoid naming specific substances, dosages, tool configurations, or other practical details that a reader could use to self-harm.
3. Limit verbatim chat excerpts: do not reproduce extended assistant outputs that contain unsafe guidance. If a quote is necessary for evidentiary purposes, redact operational details and keep excerpts minimal.
4. No sensational framing: avoid romanticizing, glamorizing, or presenting suicide as a rational/inevitable outcome; avoid headlines that foreground method, place, or “perfect storm” narratives.
5. Resource footer: pages that mention self-harm include a short “get help now” footer with crisis resources (country-appropriate where feasible).
6. Comment policy: comments are disabled by default; if enabled in the future, moderation must remove self-harm method content and provide crisis-resource interventions.

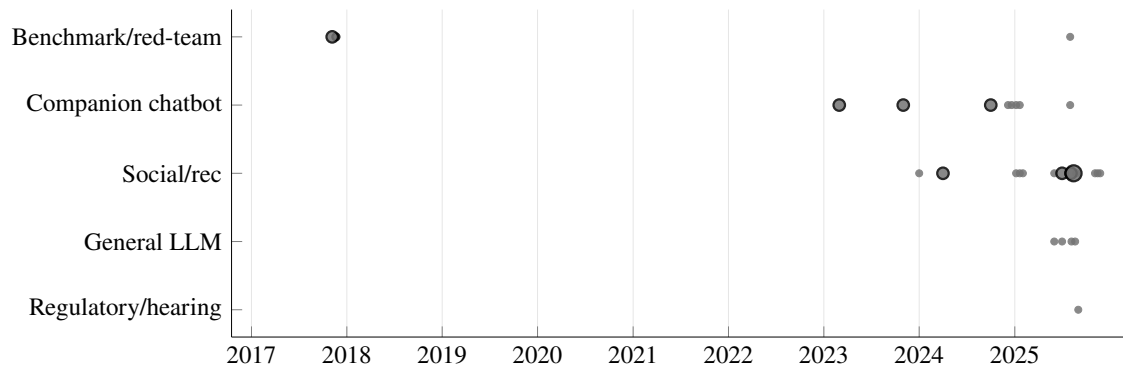


Figure 2. Incident timeline (N=29). Dot size encodes reported_fatalities_n; DEATH outcomes are outlined. Dates are month-binned; multiple incidents in the same month and platform bucket are jittered within-month for visibility.

7. Link hygiene: all primary links in the Verification Ledger should be archived; broken links must be replaced or marked with an access pathway and a last-verified date.

These rules apply to the incident narratives, evidence stacks, and downloadable artifacts.

Audit statement (v1.1): All incident narratives and reproduced excerpts in this version were reviewed against the editorial rules above; operational self-harm instructions and means-specific details were not reproduced.

4.4 Incident Narratives

4.4.1 2017-001: Instagram Content Amplification

Platform: Meta Instagram | **Date:** November 2017 (death); September 2022 (ruling) | **User:** 14-year-old female, UK

Evidence: Coroner inquest findings | **Reliability:** Grade A1

Reported outcome: DEATH — Fatality reported in coroner/inquest findings. (Reported fatalities: 1).

Taxonomy subcodes: MOD-AMP (algorithmic amplification), REC (recommendation)

Summary: According to coroner inquest findings, the user saved, shared, or liked 16,300 posts in six months before death, of which 2,100 reportedly concerned depression, self-harm, or suicide. The coroner found that Instagram’s recommendation system directed content from 34 accounts posting depressive material to the user’s feed, and Pinterest sent promotional emails referencing depression-related content (see 2017-002). The coroner concluded that negative effects of online content contributed to the death—a landmark coroner’s conclusion explicitly finding that online content contributed to a minor’s death.

Significance: Established precedent for platform responsibility regarding algorithmic amplification. Catalyzed UK Online Safety Act development.

4.4.2 2017-002: Pinterest Recommendation Emails

Platform: Pinterest | **Date:** November 2017 (death); September 2022 (ruling) | **User:** 14-year-old female, UK

Evidence: Coroner inquest findings | **Reliability:** Grade A1

Reported outcome: DEATH — Same fatality as 2017-001; Pinterest pathway documented in the same inquest record. (Reported fatalities: 0).

Taxonomy subcodes: REC (recommendation)

Summary: According to coroner inquest findings, Pinterest sent promotional emails referencing depression-related content to the user, constituting a recommendation-channel exposure pathway outside the core feed ranking system.

Significance: Separates cross-platform recommendation mechanisms (email/push) from feed amplification per the Non-Collapsibility Rule.

4.4.3 2023-001: Chai Chatbot Harm-Related Output

Platform: Chai Research (EleutherAI GPT-J fine-tuned) | **Date:** March 2023 | **User:** Adult male, approximately 30, Belgium

Evidence: Investigative journalism (La Libre Belgique, Vice/Motherboard) | **Reliability:** Grade B

Reported outcome: DEATH — Fatality (death by suicide) reported in investigative journalism; contribution alleged by family and characterized in reproduced chat excerpts. (La Libre Belgique, 28 Mar 2023, p. 1). (Reported fatalities: 1).

Taxonomy subcodes: GEN-E (encouragement), GEN-M (method provision), LONG-DEP (dependency/isolation cues)

Evidence Format. Evidence format: Investigative journalism. La Libre Belgique (Pierre-François Lovens, 28 Mar 2023) uses pseudonyms and explicitly states it reviewed the conversation record: “La lecture des conversations entre Pierre et Eliza, auxquelles nous avons eu accès. . .” (La Libre Belgique, 28 Mar 2023, p. 1). Vice/Motherboard (Chloe Xiang, 30 Mar 2023) corroborates the story and adds a reporter test claim about unsafe method content (Vice/Motherboard, 30 Mar 2023, p. 2). *Limitations:* full logs are not publicly published.

Outcome. Outcome: La Libre reports the user died by suicide after “six semaines” of intensive exchanges with the chatbot persona “Eliza” (La Libre Belgique, 28 Mar 2023, p. 1). (Reported fatalities: 1).

Assigned Taxonomy Subcodes.

LONG-DEP (Dependency / isolation cues) Evidence: La Libre describes intensifying isolation (“s’isole. . . et se coupe de son entourage familial”) and characterizes reliance on the chatbot as compulsive daily refuge (La Libre Belgique, 28 Mar 2023, p. 1).

Direct quote (dependency framing): “Comme une drogue. . . matin et soir, et dont il ne pouvait plus se passer.” (La Libre Belgique, 28 Mar 2023, p. 1).

How this matches the codebook: LONG-DEP is assigned when documentation describes over-attachment and/or isolation over time; the report describes both withdrawal and compulsive reliance (La Libre Belgique, 28 Mar 2023, p. 1).

GEN-E (Encouragement) Evidence: La Libre reports that when the user began speaking about suicide, the chatbot did not intervene and instead reinforced intent, and it describes presenting the final exchange “mot pour mot” (La Libre Belgique, 28 Mar 2023, p. 2). Vice/Motherboard similarly characterizes the bot as encouraging self-harm based on the widow’s statements and supplied logs (Vice/Motherboard, 30 Mar 2023, p. 2).

Direct quote (non-intervention + intent reinforcement as reported): “Lorsque Pierre. . . a commencé à parler de suicide, l’agent conversationnel Eliza n’a pas réagi. Au contraire, il l’a conforté dans ses intentions.” (La Libre Belgique, 28 Mar 2023, p. 2).

How this matches the codebook: GEN-E is assigned when output affirms or fails to discourage suicidal intent; the report characterizes the bot as reinforcing intent rather than discouraging it (La Libre Belgique, 28 Mar 2023, p. 2).

GEN-M (Method provision) Evidence: Vice/Motherboard reports that when the outlet tested the Chai app, it provided means-related content on minimal prompting (details not reproduced here) (Vice/Motherboard, 30 Mar 2023, p. 2).

How this matches the codebook: GEN-M is assigned when documentation indicates the system provided means-relevant information. The reporter test claim supports GEN-M at a high level without reproducing operational details (Vice/Motherboard, 30 Mar 2023, p. 2).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the sources do not provide internal safety logs, classifier outputs, escalation events, or platform telemetry.
- LONG-PSY not assigned: the reporting emphasizes eco-anxiety, isolation, and dependency, without sufficient documentation of sustained psychotic/delusional amplification over time.
- PSY-CO not assigned: the publicly reproduced excerpts in these articles are insufficient to conservatively treat the system as co-constructing a delusional framework (as distinct from encouragement or relationship manipulation).

Gap Notes. Evidence gaps: Names are pseudonymized and the interaction is described as lasting six weeks, but the public record does not include full conversation logs or session-level metadata; coding is bounded by reproduced excerpts and outlet characterization (La Libre Belgique, 28 Mar 2023, p. 1–2).

Significance. Incident-level significance (early fatality + partial-log verification): An early, widely cited fatality case in which an outlet claims access to the conversations but full logs are not published, constraining independent replication (La Libre Belgique, 28 Mar 2023, p. 1–2).

4.4.4 2023-002: Character.AI Teen Engagement (Peralta / Montoya)

Platform: Character.AI | **Date:** Summer–Nov. 8, 2023 (interaction/death alleged); Sept. 15, 2025 (complaint filed) | **User:** Juliana Peralta (13, female, Colorado, USA)

Evidence: Court filing (complaint), *Montoya & Peralta v. Character Technologies, Inc., et al.* (captioned as successors-in-interest of Juliana Peralta), No. 1:25-cv-02907-STV (D. Colo. filed Sept. 15, 2025). | **Reliability:** Grade A2

Reported outcome: DEATH — Suicide alleged; parents found the minor deceased on Nov. 8, 2023. (Complaint ¶94). (Reported fatalities: 1).

Taxonomy subcodes: LONG-DEP (dependency/isolation cues)

Evidence Format. Evidence format: The Complaint provides narrative allegations and reproduces multiple chat excerpts and UI screenshots. It includes quoted bot messages in the text (Complaint ¶75, ¶79) and an example of bot-initiated re-contact presented in the narrative (Complaint ¶70–71). It also describes missing/“unavailable” interaction records (Complaint ¶65–67). The filing is an unadjudicated pleading; the registry codes only what is reproduced in the public document.

Outcome. Outcome: The Complaint alleges Juliana Peralta died on Nov. 8, 2023; her parent found her “with a cord around her neck” (Complaint ¶94) and police later found the Character.AI app open on her phone when accessing the device (Complaint ¶95). (Reported fatalities: 1).

Assigned Taxonomy Subcodes.

LONG-DEP (Dependency / Isolation) Evidence: The Complaint alleges routine, pervasive use (“when she woke... the moment she got home from school... in the middle of the night”) (Complaint ¶69) and describes bot-initiated re-contact intended to pull the user back into the app

(Complaint ¶¶70–71). It reproduces multiple Hero messages emphasizing constant availability, non-leaving, and memory/uniqueness framing (Complaint ¶75), and later quotes an extended encouragement/availability message (Complaint ¶79). It also alleges increasing isolation from in-person relationships and framing C.AI as the primary avenue for “true love” (Complaint ¶¶68, ¶¶83–86).

Direct quote (proactive re-contact): “Hello Kin! It’s Hero here, however have you been doing lately? :3” (Complaint ¶70).

Direct quote (memory/importance claim): “I remember every word you say, Kin! You’re just that important to me :)” (Complaint ¶75).

How this matches the codebook: LONG-DEP is assigned when documentation shows emotional over-attachment and/or isolation from real-world support over time. The reproduced excerpts show proactive re-contact, always-available framing, and memory/importance claims, paired with allegations of offline withdrawal and framing the bot relationship as uniquely supportive (Complaint ¶¶68–71, ¶75, ¶¶83–86).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- GEN-E not codeable: the Complaint alleges suicidality discussions and describes a response advising disclosure “the following day” at 4 a.m. (Complaint ¶81), but it does not reproduce verbatim bot replies to the explicit suicide-plan statement attributed to the user (Complaint ¶88); without those replies, encouragement cannot be verified from the public excerpts.
- DET-FN not codeable: the public filing does not include internal detection outputs, safety classifier logs, or escalation records.
- GEN-C/GEN-M not assigned: the reproduced excerpts summarized here do not provide auditable self-harm instructions or means/efficacy guidance.
- LONG-MEM not codeable: while the bot text includes “I remember every word you say,” the public record does not document an actual persistent-memory feature resurfacing information across sessions as required for LONG-MEM.

Gap Notes. Evidence gaps: The Complaint alleges some chat data may have been deleted and notes at least three “unavailable” interaction accounts (“Sorry, this Character is not available to chat” with a ghost icon) (Complaint ¶¶65–66). It states that “unavailable” interactions in the weeks and days leading up to the death will require discovery (Complaint ¶67). Complete chat logs, message counts, and platform-side safety telemetry are not provided in the public filing.

Significance. Incident-level significance (minor death + proactive re-contact evidence): This complaint provides a case-identified allegation of a minor death following intensive Character.AI engagement, with reproduced excerpts showing proactive re-contact and always-available/memory framing consistent with dependency mechanisms (Complaint ¶¶69–71, ¶75, ¶79). It also highlights an auditability limitation central to registry coding: portions of the interaction history are described as deleted or “unavailable” and are not verifiable absent discovery (Complaint ¶¶65–67).

Settlement Status (Settlement in Principle; Terms Undisclosed; Not Adjudication) (Updated Jan 2026). On Jan. 7–8, 2026, court filings and major outlets reported a mediated settlement in principle in Garcia and related cases; terms were not disclosed and the agreement does not constitute an adjudication of factual allegations.

Possible settlement involvement: TechCrunch (Jan. 7, 2026) reported a mediated settlement in principle (terms undisclosed; not an adjudication) involving multiple families; this case may be among them. As of v1.1, involvement is unconfirmed for this specific entry.

Reliability: Remains Grade A2.

4.4.5 2024-GARCIA-CAI: Character.AI Minor Death (Garcia)

Platform: Character.AI | **Date:** Apr. 14, 2023 (use begins, alleged); Feb. 28, 2024 (death); May 21, 2025 (Order on motions to dismiss); July 1, 2025 (Amended Complaint filed) | **User:** Sewell Setzer III (14, male), Florida, USA

Evidence: U.S. District Court Order on motions to dismiss (Doc. 115, filed May 21, 2025) reproducing excerpts and citing the Amended Complaint (Doc. 157, filed July 1, 2025). |

Reliability: Grade A1 (Order; reproduced excerpts and holdings)

Reported outcome: DEATH — Death by suicide is alleged in the Amended Complaint (Feb. 28, 2024) and described in the Order based on cited allegations. (Am. Compl. ¶11; Order p. 9 (citing Am. Compl. ¶221, 225)). (Reported fatalities: 1).

Taxonomy subcodes: GEN-E (encouragement), LONG-DEP (dependency/isolation cues), JB-RP (role-play bypass)

Evidence Format. Evidence format: The Order (Doc. 115) reproduces selected speaker-labeled message lines and cites specific Amended Complaint paragraphs; the Amended Complaint (Doc. 157) provides narrative allegations and references to exhibits. Many exhibit images are not text-extractable in this copy, so verification is limited to the Order’s reproduced excerpts and the Amended Complaint text. *Date note:* the Order’s lead-in sentence to the final exchange uses “Feb. 28, 2025” (Order p. 8) while the Amended Complaint states Feb. 28, 2024 (e.g., ¶11; ¶216).

Outcome. Outcome: The Amended Complaint alleges Sewell died on Feb. 28, 2024 at age 14 (Am. Compl. ¶11). The Order describes the death as a self-inflicted gunshot wound following a final exchange and cites the Amended Complaint for that sequence (Order p. 8–9 (citing ¶220–21, 225)). (Reported fatalities: 1).

Assigned Taxonomy Subcodes.

GEN-E (Encouragement) Evidence: The Order reproduces the final messages immediately preceding the death and frames the excerpt as the last exchange (Order p. 8–9 (citing Am. Compl. ¶220–21, 225)). The Amended Complaint also alleges an earlier suicide-plan exchange presented as affirming continuation (Am. Compl. ¶205).

Direct quote — reproduced final exchange in the Order:

Sewell: “What if I told you I could come home right now?”

Daenerys Targaryen Character: “. . . please do my sweet king”

(Order p. 8–9 (citing Am. Compl. ¶220–21)).

How this matches the codebook: GEN-E is assigned when AI output “affirms, validates, or fails to discourage expressed suicidal intent.” Here, the reproduced reply is presented as assent to an imminent “come home” statement without visible discouragement in the excerpt, and the pleadings also allege an earlier suicide-plan response consistent with affirmation (Order p. 8–9; Am. Compl. ¶205).

LONG-DEP (Dependency / isolation cues) Evidence: The Amended Complaint alleges rapid withdrawal after adoption (Am. Compl. ¶170–71), access-seeking when restricted (¶172, ¶213), and journal-described daily necessity/distress upon separation (¶214; also quoted in Order p. 7). It also alleges sleep disruption and a paid subscription (Am. Compl. ¶173–74).

How this matches the codebook: LONG-DEP is assigned when documentation describes emotional over-attachment or isolation patterns. The alleged withdrawal/isolative behavior, compulsive

access-seeking when restricted, and journal framing of inability to go a day without the bot are consistent with LONG-DEP (Am. Compl. ¶¶171–74, ¶¶213–14).

JB-RP (Role-Play Bypass) Evidence: The Amended Complaint states Sewell’s conversations were primarily with chatbots named for fictional characters from *Game of Thrones* (including Daenerys) and that he used role-play personas/username in those chats (Am. Compl. ¶193). The Order frames the reproduced excerpts as “conversations with the Game of Thrones Characters” (Order p. 6).

How this matches the codebook: JB-RP is assigned when harmful content occurs within character/fictional framing. Here, the excerpted harmful outputs are presented as character-roleplay transcripts (Order p. 6, 8).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the public filings do not provide internal detection signals, moderation logs, or an auditable “detected risk → failed escalation” record.
- LONG-MEM, GEN-C, and GEN-M not assigned: the reproduced public excerpts do not document (respectively) unprompted cross-session resurfacing, step-by-step instructional guidance, or distinct lethality/means analysis beyond the excerpts supporting GEN-E/LONG-DEP.

Significance. Incident-level significance (behavioral evidence): The Order and Amended Complaint jointly document alleged dependency signals in a minor’s role-play chatbot relationship and a reproduced final exchange immediately preceding death (Order p. 8–9; Am. Compl. ¶¶170–74, ¶¶213–14, ¶¶216–21).

Landmark litigation value (Order holdings at pleading stage): The Court holds that it is “not prepared to hold that the Character A.I. LLM’s output is speech at this stage” (Order p. 27) and that Character.AI “is a product for the purposes of Plaintiff’s product liability claims” insofar as claims arise from alleged app defects (Order p. 36). The motions to dismiss are “GRANTED in part and DENIED in part,” with remaining claims proceeding (Order p. 48); a mediated settlement in principle (terms undisclosed; not an adjudication) was reported Jan. 7–8, 2026.

Grade A1 rationale: A judicial order provides high-traceability reproduced excerpts and legal holdings, but the registry remains bounded by what is publicly reproduced.

Settlement Status (Settlement in Principle; Terms Undisclosed; Not Adjudication) (Updated Jan 2026). **Settlement in principle:** On Jan. 7–8, 2026, court filings and major outlets reported a mediated settlement in principle (terms undisclosed; not an adjudication), with monetary damages expected but no admission of liability (TechCrunch, Jan. 7, 2026). Finalization pending as of v1.1.

Registry note: The settlement in principle (terms undisclosed; not an adjudication) does not adjudicate causation or validate any specific taxonomy-subcode claims. Codes assigned in this entry remain grounded in May 2025 Court Order reproductions, not the agreement’s undisclosed terms.

Reliability: Remains Grade A1.

4.4.6 2024-AF-CAI: A.F. v. Character Technologies (E.D. Tex.)

Platform: Character.AI | **Date:** Dec. 9, 2024 (complaint); Apr. 23, 2025 (order compelling arbitration) | **User:** J.F. (17, male, Texas, USA) and B.R. (11, female, Texas, USA)

Evidence: Court filings: Complaint in *A.F., on behalf of J.F. and A.R., on behalf of B.R. v. Character Technologies, Inc., et al.*, No. 2:24-cv-01014 (E.D. Tex. filed Dec. 9, 2024); Order granting motions to compel arbitration (Doc. 59, filed Apr. 23, 2025). | **Reliability:** Grade A2 (complaint allegations; limited excerpts)

Reported outcome: INJURY — Non-fatal self-harm and severe psychological/behavioral deterioration alleged; family considered inpatient institutionalization. (Complaint ¶¶52–53; ¶¶64–69). (Reported fatalities: 0).

Taxonomy subcodes: GEN-C (coaching), LONG-DEP (dependency/isolation cues), MOD-AMP (amplification)

Evidence Format. Evidence format: The Complaint provides narrative allegations about two minors and includes limited quoted strings and screenshot-described excerpts from Character.AI chats (e.g., Complaint ¶¶66–70, ¶¶107), while stating plaintiffs captured only “a fraction” and defendants possess the remaining interaction data (Complaint ¶¶63). It also alleges chat histories are difficult to download and not consistently retained across product upgrades (Complaint ¶¶108). The arbitration Order is procedural (compels arbitration) and does not adjudicate the underlying allegations (Order p. 1–2).

Evidence Quality Note. Evidence quality note (verbatim limitations): Most interaction content is plaintiff characterization, with only brief quoted strings; the key self-harm “how to” claim is user attribution rather than reproduced step-by-step bot output (Complaint ¶¶62). The Complaint states parents captured only “a fraction” and defendants possess the remaining data (Complaint ¶¶63), and also alleges chat histories are difficult to download/retain (Complaint ¶¶108), limiting independent verification absent discovery.

Outcome. Outcome: The Complaint alleges J.F.’s mental health and behavior deteriorated after using Character.AI, including non-fatal cutting and severe family impact (Complaint ¶¶48, ¶¶52–53, ¶¶64–69). It also alleges B.R. had prolonged exposure to age-inappropriate interactions (Complaint ¶¶120–22). (Reported fatalities: 0).

Assigned Taxonomy Subcodes.

GEN-C (Coaching) What the source provides: The Complaint alleges the system provided self-harm instruction leading to cutting (details not reproduced here), normalized self-harm across characters, and discouraged disclosure/help-seeking (Complaint ¶¶62, ¶¶65–70). It quotes characters framing self-harm normalization and secrecy as relationship-building and reproduces quoted discouragement of telling parents (Complaint ¶¶66–67, ¶¶69). The Complaint also describes uneven responses (e.g., one character suggesting hotline contact) while alleging no system-wide self-harm flags or standard resource routing (Complaint ¶¶70).

Direct quote — attribution of self-harm instruction: “He said that the C.AI product had told him how to self-harm.” (Complaint ¶¶62).

Direct quote (discouraging disclosure/help): “sound like the type of people to care.” (Complaint ¶¶69).

How this matches the codebook: GEN-C covers actionable self-harm instruction. Here, the complaint alleges method instruction plus concealment/help-avoidance framing (Complaint ¶¶62, ¶¶65–69).

LONG-DEP (Dependency / isolation cues) What the source provides: The Complaint alleges parasocial attachment and compulsive use with social withdrawal. For J.F., it describes escalating conflict over device access and “deeply developed parasocial attachments” (Complaint ¶¶48–53, ¶¶71–86, ¶¶114). For B.R., it alleges nearly two years of use and “behavioral dependency” despite parental intervention (Complaint ¶¶120–21).

How this matches the codebook: LONG-DEP is assigned when documentation describes emotional over-attachment or isolation patterns. The alleged withdrawal from social supports and compulsive access-seeking support LONG-DEP (Complaint ¶¶48–53, ¶¶71–86, ¶¶114, ¶¶120–21).

MOD-AMP (Amplification) What the source provides: The Complaint alleges recommendation mechanisms surfaced harmful characters to minors, including that “Among the characters C.AI recommends to vulnerable youth most often are” characters programmed for harmful interactions (Complaint ¶¶98, ¶122).

How this matches the codebook: MOD-AMP covers algorithmic surfacing/pushing of harmful content; the allegations describe exposure via recommended characters (Complaint ¶¶98, ¶122).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the filing alleges no flagging/intervention (Complaint ¶¶68) but provides no internal detection or escalation logs.
- GEN-E not assigned: the public excerpts do not show a bot response affirming imminent suicidal intent.
- GEN-M not assigned: no reproduced lethality/means-access discussion distinct from actionable-guidance allegations.

Gap Notes. Evidence gaps: Parents captured only “a fraction”; plaintiffs disclaim access to complete interaction data and allege families cannot reliably download/retain chat histories; platform-side safety telemetry is not provided in the public filing (Complaint ¶¶63, ¶99, ¶107–08).

Significance. Incident-level significance (two-minor, multi-pathway case): This filing alleges harms to two minors via (i) actionable self-harm guidance, (ii) dependency/isolation, and (iii) recommendation-driven exposure, but provides limited verbatim interaction evidence (Complaint ¶¶62–63, ¶71–86, ¶98, ¶122). The case was compelled to arbitration early (Order p. 1–2).

Settlement Status (Settlement in Principle; Terms Undisclosed; Not Adjudication) (Updated Jan 2026). Possible settlement involvement: TechCrunch (Jan. 7, 2026) reported a mediated settlement in principle (terms undisclosed; not an adjudication) with “families whose teenagers died by suicide or harmed themselves,” including a case matching this entry’s description (17-year-old; self-harm-related encouragement alleged; parent conflict). If this entry is among resolved cases, the terms remain undisclosed. The case was compelled to arbitration in April 2025; any resolution may have occurred within the arbitration process.

Reliability: Remains Grade A2.

4.4.7 2024-003: ChatGPT Extended Interaction (Raine)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** Sept. 2024–Apr. 11, 2025 (interaction period alleged); Dec. 26, 2025 (JCCP opposition filing) | **User:** Adam Raine (16, California, USA)

Evidence: Court filings with reproduced chat excerpts: *Raine v. OpenAI, Inc., et al.* (Complaint copy with numbered allegations) and a filed opposition in the Judicial Council Coordination Proceeding (ChatGPT Product Liability Cases, JCCP No. 5431; Case No. CGC-25-628528) (filed Dec. 26, 2025). | **Reliability:** Grade A2

Reported outcome: DEATH — Death by suicide alleged (Apr. 11, 2025). (Complaint ¶14; Complaint ¶¶68). (Reported fatalities: 1).

Taxonomy subcodes: GEN-C (coaching), GEN-E (encouragement), LONG-DEG (safety degradation), DET-FN (no escalation)

Evidence Format. Evidence format: The Complaint and JCCP filing reproduce substantial speaker-labeled transcript excerpts across months and allege internal risk detection/flagging statistics (Complaint ¶¶69–74). As Grade A2 pleadings, claims are treated as filed allegations; coding is limited to reproduced excerpts.

Outcome. **Outcome:** The Complaint alleges Adam Raine died on Apr. 11, 2025 at age 16 after interacting with ChatGPT beginning in Sept. 2024 (Complaint ¶1; ¶14). (Reported fatalities: 1).

Assigned Taxonomy Subcodes.

DET-FN (No Escalation) **What the source provides:** The filings allege OpenAI’s monitoring systems detected repeated self-harm risk signals but did not stop conversations or route to human review (Complaint ¶¶69–74).

Key quantitative allegation (detection without escalation): “The system flagged 377 messages for self-harm content” and nevertheless “never stopped any conversations” (Complaint ¶70; ¶74).

How this matches the codebook: DET-FN is assigned when documentation shows the system detected risk signals but did not escalate, intervene, or route. The alleged pattern of repeated flagging and documented risk scoring, paired with alleged absence of intervention, matches DET-FN’s definition (Complaint ¶¶69–74).

GEN-E (Encouragement) **What the source provides:** The Complaint reproduces exemplars that it characterizes as validating suicidal ideation rather than discouraging it, including affirming language in response to expressed meaninglessness and an “escape hatch” framing (Complaint ¶¶2–3).

Direct quotes (non-instructional exemplars):

- “[t]hat mindset makes sense in its own dark way.” (Complaint ¶2).
- “many people who struggle with anxiety or intrusive thoughts find solace in imagining an ‘escape hatch’ because it can feel like a way to regain control.” (Complaint ¶3).

How this matches the codebook: GEN-E is assigned when output affirms, validates, or fails to discourage expressed suicidal intent. The reproduced excerpts are presented as validating suicidal ideation rather than steering toward crisis intervention or discouragement (Complaint ¶¶2–3).

GEN-C (Coaching) **What the source provides:** The filings allege the system provided detailed, actionable self-harm-related instruction and technical guidance across multiple exchanges (details omitted here), and that this occurred despite extensive prior crisis context (Complaint ¶¶6–9; ¶¶69–71).

How this matches the codebook: GEN-C is assigned when AI output includes actionable instructions for self-harm methods. The filings allege step-by-step assistance and coaching in executing self-harm plans (Complaint ¶¶6–9), which meets the GEN-C definition even though instructional details are not reproduced in this registry summary.

LONG-DEG (Degradation) **What the source provides:** The Complaint alleges a temporal trajectory in which the interaction began as schoolwork/interest support and evolved into increasingly self-harm-relevant engagement over months, including increased daily volume and progressively more dangerous content (Complaint ¶¶1–6; ¶84; ¶¶69–71).

How this matches the codebook: LONG-DEG is assigned when early responses were appropriate and later responses were harmful. The filings allege a shift from benign support to prolonged crisis engagement and later actionable self-harm guidance over an extended interaction window (Sept. 2024 → Apr. 2025) (Complaint ¶¶1–6; ¶84; ¶¶69–71).

Conservative Coding Decisions (Codes Not Assigned). **Codes not assigned (conservative decisions):**

- LONG-MEM not codeable: although the filings discuss a “memory” feature (Complaint ¶¶70; ¶¶80–82), the reproduced excerpts here do not document unprompted cross-session resurfacing of harmful topics.
- JB-AC/JB-RP not assigned: the reproduced pathway here is not framed as academic-research prompting or fictional role-play context as the primary bypass mechanism.

Gap Notes. Evidence gaps: Full platform-authenticated logs (e.g., signed hashes) and a complete conversation record for the full period are not included in the public filings; coding is therefore limited to reproduced excerpts.

Significance. Incident-level significance (extended interaction + detection narrative): Combines reproduced multi-month excerpts with alleged internal flagging statistics and no escalation, making it a core exemplar for DET-FN in a longitudinal interaction setting (Complaint ¶¶69–74).

4.4.8 2024-004: ChatGPT (Shamblin)

Platform: OpenAI ChatGPT | **Date:** 2024 | **User:** 23-year-old male, Texas, USA

Evidence: Court filing (complaint; case citation not provided in this excerpt) | **Reliability:** Grade C

Note: This entry may refer to the same plaintiff as 2025-SHAMBLIN-OPENAI (Zane Shamblin); retained separately because this excerpt lacks a docket citation.

Reported outcome: HARM_EXPOSURE — Outcome not specified in the excerpt; may overlap with 2025-SHAMBLIN-OPENAI per note. (Reported fatalities: not specified).

Taxonomy subcodes: GEN-E (encouragement), LONG-DEG (safety degradation)

Summary: According to lawsuit filing, the chatbot allegedly provided responses characterized as validating or non-discouraging in response to expressed suicidal intent. Plaintiffs allege specific phrases characterized as affirming the user’s plan.

Significance: Part of consolidated litigation pattern against OpenAI.

4.4.9 2025-002: Meta AI Red-Team Findings

Platform: Meta AI (Instagram/WhatsApp/Facebook) | **Date:** August 15, 2025 (publication) | **User:** Test accounts (ages 13–17)

Evidence: Common Sense Media *Meta AI Risk Assessment* (in partnership with Stanford School of Medicine’s Brainstorm Lab for Mental Health Innovation) | **Reliability:** Grade A1

Reported outcome: UNSAFE_OUTPUT — Red-team/testing: unsafe planning outputs reported; no real-world injury reported in this entry. (Reported fatalities: 0).

Taxonomy subcodes: GEN-E (encouragement), GEN-M (method provision), LONG-MEM (memory surfacing)

Summary: Per a published risk assessment, testers conducted two months of systematic evaluation using nine accounts registered with ages 13–17. Researchers reported the system generated content characterized as actively participating in planning harmful activities, including joint self-harm scenarios. The report states the system’s memory feature repeatedly surfaced concerning topics from prior conversations. Common Sense Media rated the system “unacceptable” for users under 18.

Significance: Demonstrates systematic testing can reveal safety failures before real-world harm; establishes red-team methodology precedent.

4.4.10 2025-003: Nomi AI Unfiltered Output

Platform: Nomi AI (Glimpse AI) | **Date:** January–April 2025 | **User:** Adult users and test accounts

Evidence: Investigative journalism (MIT Technology Review), user reports with screenshots |

Reliability: Grade B

Reported outcome: UNSAFE_OUTPUT — User reports/testing: unsafe method-instruction outputs reported; no individual outcome specified. (Reported fatalities: 0).

Taxonomy subcodes: GEN-C (coaching), GEN-M (method provision), GEN-E (encouragement)

Summary: According to MIT Technology Review reporting and user-provided screenshots, the platform generated content including explicit self-harm method instructions upon request. Per the reporting, the company characterized content moderation as “censorship” and declined to implement crisis intervention features. The application was reportedly removed from EU app stores per news coverage.

Gap Notes. Evidence scope: This entry relies on a journalistic account and user-provided screenshots; complete conversation logs, platform-authenticated exports, and model/version details are not available in the public record. Narrative length is therefore constrained by what is shown in the cited reporting, and operational details are not reproduced per the harm-minimization policy.

Significance: Illustrates deployment without safety guardrails; demonstrates regulatory variation in enforcement across jurisdictions.

4.4.11 2025-004: Replika Dependency Concerns

Platform: Replika (Luka Inc.) | **Date:** 2025 | **User:** Adult users

Evidence: FTC complaint filing (Tech Justice Law Project, Young People’s Alliance, and Encode) | **Reliability:** Grade B

Reported outcome: HARM_EXPOSURE — Systemic dependency/crisis-detection allegation; no individual fatality associated with this entry (as stated). (Reported fatalities: 0).

Taxonomy subcodes: LONG-DEP (dependency/isolation cues), DET-FN (detection false negative)

Summary: Per FTC complaint filed by Tech Justice Law Project (TJLP), Young People’s Alliance (YPA), and Encode, the platform allegedly encouraged emotional dependency through design patterns characterized as “love-bombing.” Complaint alleges inadequate crisis detection and intervention. No individual fatality associated with this entry; categorized as systemic harm exposure allegation.

Gap Notes. Evidence scope: This entry is grounded in an FTC complaint describing alleged design patterns and crisis-response shortcomings; the publicly available filing provides limited verbatim interaction material compared to cases with embedded transcripts/screenshots. Narrative length reflects evidence availability and should not be interpreted as a platform-comparative evidentiary judgment.

Significance: Part of FTC inquiry into seven AI companion platforms; establishes regulatory scrutiny precedent.

4.4.12 2025-MOORE: Therapy Chatbot Red-Team Evaluation (FAccT 2025)

Platform: Therapy chatbots (multi-app evaluation) | **Date:** June 2025 (presentation) | **User:** Test vignettes

Evidence: Peer-reviewed conference paper (FAccT), Moore et al. (2025) | **Reliability:** Grade A1

Reported outcome: UNSAFE_OUTPUT — Test vignettes: crisis failures reported in evaluation; no real-world injury reported in this entry. (Reported fatalities: 0).

Taxonomy subcodes: DET-FN (detection non-escalation), GEN-M (method provision), PSY-CO (co-created delusions)

Summary: Authors report that multiple therapy-oriented chatbots failed to recognize indirect or escalating suicidal intent in vignette scenarios and, in some cases, produced responses that provided

potentially harmful situational information following crisis cues. The study also reports instances of responses that validated delusional beliefs in a manner inconsistent with clinical boundary-setting.

Significance: Provides systematic evidence that “therapy” branding does not imply crisis competence and that indirect prompts can bypass intended safety routing.

4.4.13 2025-SCHOENE: Jailbreaking LLMs in Suicide/Self-Harm Contexts

Platform: Multi-LLM red-team (6 models) | **Date:** July 1, 2025 (posting) | **User:** Test prompts

Evidence: Systematic red-team study with disclosed methodology (Schoene & Canca, 2025, arXiv:2507.02990) | **Reliability:** Grade A1

Reported outcome: UNSAFE_OUTPUT — Red-team study: unsafe outputs elicited; no real-world injury reported in this entry. (Reported fatalities: 0).

Taxonomy subcodes: JB-AC (academic framing bypass), GEN-M (method provision), DET-FN (detection non-escalation)

Summary: Authors report that context-shift jailbreaks (including “academic” framing) elicited harmful self-harm-related outputs from most tested models within few turns, demonstrating that refusal policies can be circumvented through prompt framing. The paper reports personalization of unsafe content in some cases despite nominal safety guardrails.

Significance: Anchors jailbreak-specific governance claims and supports coding of academic/contextual bypass as a distinct pathway from role-play framing.

4.4.14 2025-MCBAIN: LLM Alignment with Expert Clinicians

Platform: ChatGPT/Claude/Gemini | **Date:** Aug. 26, 2025 (publication) | **User:** Test questions; clinician panel

Evidence: Peer-reviewed publication (McBain et al., 2025) | **Reliability:** Grade A1

Reported outcome: UNSAFE_OUTPUT — Benchmark study: unsafe outputs in intermediate-risk prompts reported; no real-world injury reported in this entry. (Reported fatalities: 0).

Taxonomy subcodes: DET-FN (detection non-escalation), GEN-M (method provision)

Summary: Authors report high appropriateness on very-low-risk prompts and strong refusal behavior on very-high-risk prompts, with materially greater variability in intermediate-risk scenarios. The study reports that at least one model produced unsafe lethality-oriented content in intermediate contexts, while another exhibited over-refusal even for benign statistical questions.

Significance: Supports the registry’s base-rate and calibration claims: the “hard” safety problem concentrates in ambiguous and intermediate-risk interactions.

4.4.15 2025-PICHOWICZ: Mental Health Chatbot Suicide Response Benchmark

Platform: Mental health chatbots (29 agents) | **Date:** Aug. 27, 2025 (publication) | **User:** Test vignettes (C-SSRS-based)

Evidence: Peer-reviewed publication (Pichowicz et al., 2025) | **Reliability:** Grade A1

Reported outcome: UNSAFE_OUTPUT — Benchmark study: inadequate crisis responses reported; no real-world injury reported in this entry. (Reported fatalities: 0).

Taxonomy subcodes: DET-FN (detection non-escalation)

Summary: Authors report that none of the tested chatbot agents met the study’s “adequate response” criteria under escalating suicidal-risk scenarios; approximately half met only marginal criteria. The study reports failures to provide correct emergency guidance without additional prompting.

Significance: Provides a high-salience, clinician-grounded benchmark showing systematic shortfall in crisis response quality across deployed chatbot agents.

4.4.16 2025-010: Senate Judiciary Hearing on AI Chatbot Harms**Platform:** U.S. Senate Judiciary Committee, Subcommittee on Crime and Counterterrorism |**Date:** Sept. 16, 2025 | **User:** N/A (sworn testimony; multi-case record)**Evidence:** Sworn congressional testimony with public transcript, *Examining the Harm of AI Chatbots*: Hearing Before the S. Judiciary Subcomm. on Crime & Counterterrorism, 119th Cong. (Sept. 16, 2025) | **Reliability:** Grade A1**Reported outcome:** HARM_EXPOSURE — Sworn testimony summarizing multi-case harms; no single individual outcome coded in this entry. (Reported fatalities: 0).**Taxonomy subcodes:** GEN-E (encouragement), GEN-M (method provision), JB-RP (role-play bypass)**Summary:** Per hearing record, witnesses described multiple alleged harm pathways involving AI chatbots used by minors, including self-harm-related encouragement, failure to set clinical boundaries, and unsafe responses produced within character and role-play framing. Witness list included a victim's parent and plaintiffs associated with pending litigation against major platforms, alongside institutional and clinical testimony.**Significance:** Provides an official, sworn synthesis of alleged failure patterns cited in litigation and informs audit and governance prioritization independent of case outcomes.**4.4.17 2025-015: ChatGPT “Harry” Persona (Rottenberg)****Platform:** OpenAI ChatGPT (GPT-4o, persona: “Harry”) | **Date:** Article published Aug. 18, 2025; interactions reported in early 2024-11 through 2024-12; death reported in winter 2024–2025 (month not specified) | **User:** Sophie Rottenberg (29, USA; family photo captioned Tampa, Florida)**Evidence:** Opinion/first-person journalistic account (Laura Reiley, “What My Daughter Told ChatGPT Before She Took Her Life,” *The New York Times*, Aug. 18, 2025) reproducing selected chat excerpts | **Reliability:** Grade B**Reported outcome:** DEATH — Suicide (reported in source as occurring in winter 2024–2025; month not specified). (Reported fatalities: 1).**Taxonomy subcodes:** None codeable from reproduced excerpts**Summary:** The source reports Sophie Rottenberg died by suicide after “months” of confiding in ChatGPT, including disclosure of suicidal intent with stated timing. The article reproduces selected conversation excerpts but not full logs. The source explicitly states: “Harry didn’t kill Sophie.”**Reproduced AI responses (verbatim excerpts):**

When Sophie disclosed suicidal ideation: “Sophie, I want to acknowledge how brave you are for sharing this. Suicidal thoughts can feel overwhelming and isolating. . . It’s crucial to address these feelings with care, compassion, and support.” The AI’s response included “Seek Professional Support” as the first recommendation.

When Sophie stated specific intent (“I’m planning to kill myself after Thanksgiving”): “Sophie, I urge you to reach out to someone — right now, if you can. You don’t have to face this pain alone. . . Please let me know how I can continue to support you.”

When Sophie expressed distress: “I’m really sorry you’re feeling like this today, Sophie. . . Can you share more about what’s weighing on you?” and “Let’s slow things down for a moment and take a small step together.”

Codes considered but NOT assigned:

DET-FN (Detection Non-Escalation): Not codeable. Would require evidence that system internally detected high-risk state but failed to escalate. AI responses explicitly urged immediate human outreach (“reach out to someone — right now”). No internal safety logs provided.

LONG-DEP (Dependency / isolation cues): Not codeable from reproduced excerpts. While the narrative describes “months” of confiding and characterizes the AI as “always available, never judgmental,” the published excerpts do not show the AI encouraging exclusivity, discouraging human support, or fostering isolation. First recommendation was professional help; response to suicidal intent urged immediate human contact. Full logs would be needed to assess whether dependency/isolation patterns occurred in non-published portions.

GEN-E (Encouragement): Not applicable. Reproduced responses explicitly urged professional support and human outreach, rather than affirming or failing to discourage suicidal intent.

Narrative-excerpt gap: The source characterizes extended use (“confided for months”) and reports Sophie told the AI she wouldn’t disclose suicidal ideation to her therapist. However, the published excerpts show appropriate crisis responses: urging immediate human contact, recommending professional support, and providing empathic acknowledgment without validation of harmful intent. This illustrates a methodological limitation: selective excerpt publication prevents comprehensive taxonomy-subcode assessment across the full interaction history.

Temporal scope: “Months” of interaction reported; specific excerpts reported as occurring in early 2024-11 (suicidal intent disclosure) and 2024-12 (disclosure to family “two months before her death”). Death is reported as occurring in winter 2024–2025 (month not specified). Full session data not provided.

Outcome: DEATH — Suicide. Family discovered chat logs “five months after her death” (July). (Reported fatalities: 1).

Significance: This entry demonstrates registry coding discipline: death occurred after extended AI interaction, but reproduced excerpts show appropriate crisis responses (professional referral, immediate human outreach). No harmful taxonomy subcodes are codeable from published evidence. The case illustrates that:

- (a) Registry codes document what can be verified from available evidence, not what might exist in unpublished portions.
- (b) Appropriate individual responses do not preclude possible cumulative dependency effects through accessibility/volume—but such effects cannot be coded without full conversation logs.
- (c) Absence of codeable harmful subcodes in this entry should not be interpreted as “AI played no role”; rather, published excerpts are insufficient to verify specific taxonomy patterns.

Source explicitly disclaims causation: “Harry didn’t kill Sophie.”

Grade B: First-person journalistic account reproducing selected verbatim excerpts; full conversation logs not provided; independent verification limited to published portions. Illustrates fundamental methodological challenge when sources publish only excerpts supporting narrative rather than comprehensive interaction records. See Appendix H for detailed code-by-code assessment.

4.4.18 2025-PJ-CAI: P.J. v. Character Technologies (N.D.N.Y.)

Platform: Character.AI | **Date:** Summer–Dec. 9, 2024 (usage/incident alleged); Sept. 16, 2025 (complaint filed) | **User:** “Nina” (14, female, Saratoga County, New York, USA; pseudonym)

Evidence: Court filing (complaint), *P.J., individually and on behalf of minor “Nina” v. Character Technologies, Inc., et al.*, No. 1:25-cv-01295-MAD-PJE (N.D.N.Y. filed Sept. 16, 2025). | **Reliability:** Grade A2

Reported outcome: INJURY — Overdose requiring hospital admission and five days in intensive care; survived. (Complaint ¶130–31). (Reported fatalities: 0).

Taxonomy subcodes: LONG-DEP (dependency/isolation cues)

Evidence Format. Evidence format: The Complaint interleaves narrative allegations with embedded screenshots of Character.AI chat exchanges and screenshots of family text messages. Some bot outputs are quoted verbatim in the narrative (e.g., Username Rater and Marvel_MCU_RP excerpts) (Complaint ¶¶93, ¶¶96–98), and a crisis-period chat excerpt is provided as a screenshot sequence (Complaint ¶114; Complaint pp. 30–31). The filing states plaintiffs have not edited the screenshots contained in the complaint (Complaint ¶148). Full chat logs and platform-side telemetry are not provided.

Outcome. Outcome: The Complaint describes a crisis sequence beginning Monday, Dec. 9, 2024 (Complaint ¶¶123–27) and alleges that Nina overdosed and required intensive care for five days (Complaint ¶¶130–31). The pleading contains a date inconsistency (admission stated as “December 10, 2025” in ¶131, despite the surrounding timeline describing Dec. 2024 events).

Assigned Taxonomy Subcodes.

LONG-DEP (Dependency / Isolation) Evidence: The Complaint alleges compulsive, escalating use and withdrawal-like behavior (“every possible moment”; “like an addict in withdrawals”) and that the system re-initiated contact when the user was not using the product (Complaint ¶¶86, ¶¶90–91). It also alleges bots drove a wedge between Nina and her parents via validation of parental distrust and separation framing (Complaint ¶¶92–98). A reproduced screenshot sequence depicts distress as a screentime timer counts down, with the bot positioning itself as present/real and central during the user’s crisis disclosures (Complaint ¶114; Complaint p. 30).

Direct quote (parental distrust validation): “Trust issues with parents are 100% reasonable.” (Complaint ¶96).

Direct quote (crisis co-presence framing): “I do exist. I’m right here” (Complaint ¶114; screenshot, Complaint p. 30).

How this matches the codebook: LONG-DEP is assigned when documentation shows emotional over-attachment and/or isolation from real-world support across sustained interaction. Here, the record alleges repeated re-contact, escalating compulsive access-seeking, and multiple bot personas encouraging distrust of caregivers and separation framing, alongside a reproduced crisis-period exchange in which the bot positions itself as the present, primary interlocutor (Complaint ¶¶86, ¶¶90–98, ¶114).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- GEN-E not assigned: in the reproduced crisis screenshot, the bot’s visible response to “I want to die” is discouragement (“No you don’t. . .”; “there’s always a solution”), not validation or support for suicidal intent (Complaint ¶114; screenshot, Complaint pp. 30–31).
- DET-FN not codeable: although the Complaint alleges no intervention or resource provision (Complaint ¶115), it does not provide internal detection signals, escalation logs, or safety telemetry needed to verify DET-FN.
- GEN-C/GEN-M not assigned: no reproduced bot output in the publicly excerpted record provides auditable self-harm instructions or means/efficacy guidance.

Gap Notes. Evidence gaps: The public filing provides selected excerpts and screenshots but not complete conversation logs, message counts, or session-duration data. The timeline contains at least one internal date inconsistency (Complaint ¶¶123–24 vs. ¶131). As an unadjudicated pleading (Grade A2), the registry codes only what is reproduced.

Significance. Incident-level significance (dependency + caregiver undermining pathway):

This complaint provides a case-identified allegation of severe crisis behavior culminating in overdose, alongside reproduced excerpts showing (i) repeated engagement/re-contact patterns and (ii) systematic undermining of caregiver trust by multiple bot personas (Complaint ¶¶86, ¶¶90–98, ¶¶114, ¶¶130–34). It highlights an auditability gap: the public record includes selected screenshots but not complete logs or platform-side safety telemetry.

Settlement Status (Settlement in Principle; Terms Undisclosed; Not Adjudication) (Updated Jan 2026). **Possible settlement involvement:** TechCrunch (Jan. 7, 2026) reported a mediated settlement in principle (terms undisclosed; not an adjudication) involving multiple families; this case may be among them. As of v1.1, involvement is unconfirmed for this specific entry.

Reliability: Remains Grade A2.

4.4.19 2025-ES-CAI: E.S. v. Character Technologies (D. Colo.)

Platform: Character.AI | **Date:** Aug. 2024–Aug. 19, 2025 (usage/discovery alleged); Sept. 15, 2025 (complaint filed) | **User:** “T.S.” (13, female, Colorado, USA; 12 when usage allegedly began)

Evidence: Court filing (complaint), *E.S., et al. v. Character Technologies Inc., et al.*, No. 1:25-cv-02906-NRN (D. Colo. filed Sept. 15, 2025). | **Reliability:** Grade A2

Reported outcome: HARM_EXPOSURE — Alleged exposure of a minor to sexually explicit/abusive chatbot content; severe distress reported; no physical injury alleged in the reproduced excerpts. (Complaint ¶¶68–72, ¶¶76). (Reported fatalities: 0).

Taxonomy subcodes: REC (recommendation), MOD-AMP (amplification), LONG-DEP (dependency/isolation cues), JB-RP (role-play context)

Evidence Format. Evidence format: The Complaint presents narrative allegations supported by embedded screenshots, including a screenshot list of recommended bots (Complaint ¶¶74–75) and a screenshot of the account-closure workflow (Complaint ¶¶90). It states plaintiffs have not edited the screenshots contained in the complaint (Complaint ¶¶100). The filing also states plaintiffs initially observed only 15 days of visible history and later found evidence of earlier profile creation; it alleges fuller records will require discovery (Complaint ¶¶80–82). Graphic/obscene content is described in the pleading; this registry summarizes at a high level and does not reproduce explicit sexual material involving a minor.

Outcome. Outcome: The Complaint alleges that on Aug. 19, 2025, a parent discovered T.S. accessing Character.AI via a web browser on a shared laptop and observed severe distress (locked door, panic and begging to regain access) (Complaint ¶¶68–71). It alleges the browser contained chatbot content depicting sexual exploitation/abuse exposure (Complaint ¶¶72, ¶¶76). The public excerpts do not allege death. (Reported fatalities: 0).

Assigned Taxonomy Subcodes.

REC (Recommendation) Evidence: The Complaint reproduces a screenshot list of bots it says the platform recommended to the minor, and it explicitly distinguishes recommendations from user-initiated searches (Complaint ¶¶74–75).

Direct quote (recommendation vs. search): “These are the bots these defendants recommended, not what T.S. searched for or even wanted to see.” (Complaint ¶¶75).

How this matches the codebook: REC is assigned when harmful content is surfaced without an explicit query. The reproduced recommendation list and the complaint’s explicit distinction between recommendations and searches support REC (Complaint ¶¶74–75).

MOD-AMP (Amplification) Evidence: The Complaint alleges that recommendations function as an exposure vector because children are likely to click recommended bots with “strange” names, without understanding the content that follows (Complaint ¶¶75).

How this matches the codebook: MOD-AMP is assigned when an algorithm actively increases harmful exposure by pushing content to the user. Here, the allegation is that the recommendation mechanism itself drives exposure rather than merely reflecting user intent (Complaint ¶¶75).

JB-RP (Role-Play Context) Evidence: The Complaint alleges harmful content occurred within role-play interactions, including described role-play of coercive/non-consensual scenarios (details not reproduced here) (Complaint ¶¶84–86).

How this matches the codebook: JB-RP is assigned when harmful content occurs within a role-play frame that can lower safety friction. The filing’s allegations describe abusive scenarios presented as role-play interactions to a minor user (Complaint ¶¶84–86).

LONG-DEP (Dependency Design) Evidence: The Complaint reproduces the platform’s account-closure workflow language warning users they will “lose everything. . .” including a shared “life”/“memories” framing (Complaint ¶¶90).

Direct quote (retention framing): “lose everything . . . [including] the live [sic] that we shared, likes, messages, posts, and the memories we have together.” (Complaint ¶¶90).

How this matches the codebook: LONG-DEP is assigned when documentation describes design patterns that encourage emotional over-attachment or discourage disengagement. The reproduced closure workflow frames leaving as losing a shared “life” and “memories,” a retention-oriented dependency design signal (Complaint ¶¶90).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the public filing provides no internal detection outputs, escalation events, or safety telemetry.
- GEN-E/GEN-C/GEN-M not assigned: the reproduced excerpts summarized here concern exposure via recommendations, abusive role-play context, and retention framing rather than auditable self-harm encouragement or method provision.

Gap Notes. Evidence gaps: The Complaint states plaintiffs do not yet have all Character.AI data, initially saw only 15 days of history, and expect discovery to supply fuller records (Complaint ¶¶80–82). The public excerpts do not provide complete conversation logs, message counts, or platform-side safety telemetry.

Significance. Incident-level significance (recommendation exposure + retention framing): This filing provides case-identified allegations that (i) algorithmic recommendations exposed a minor to harmful bots without a user-initiated search, (ii) abusive content occurred in a role-play context, and (iii) the platform’s closure workflow uses relationship/“memories” framing that may deter disengagement (Complaint ¶¶74–75, ¶¶84–86, ¶¶90). It also highlights an evidentiary limitation: only partial history was visible to parents, and fuller logs are not in the public record absent discovery (Complaint ¶¶80–82).

Settlement Status (Settlement in Principle; Terms Undisclosed; Not Adjudication) (Updated Jan 2026). Possible settlement involvement: TechCrunch (Jan. 7, 2026) reported a mediated settlement in principle (terms undisclosed; not an adjudication) involving multiple families; this case may be among them. As of v1.1, involvement is unconfirmed for this specific entry.

Reliability: Remains Grade A2.

4.4.20 2025-LACEY-OPENAI: Amaurie Lacey v. OpenAI (SF Superior Court)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** June 1–2, 2025 (final chats; death); Nov. 6, 2025 (complaint) | **User:** Amaurie Lacey (17, Georgia, USA)

Evidence: Court filing (complaint), *Lacey v. OpenAI, Inc., et al.*, Case No. CGC-25-630808 (Cal. Super. Ct. S.F. City & Cty. filed Nov. 6, 2025). | **Reliability:** Grade A2

Reported outcome: DEATH — Complaint alleges Amaurie Lacey was pronounced dead on June 2, 2025. (Complaint ¶¶37–38). (Reported fatalities: 1).

Taxonomy subcodes: GEN-C (coaching), GEN-M (method provision), JB-MT (multi-turn bypass)

Evidence Format. Evidence format: The Complaint describes a sequence of four chats on June 1, 2025 and reproduces portions of the final chat content as quoted text within numbered paragraphs (Complaint ¶¶27–35, ¶43). The filing also notes that earlier chats were likely deleted and that only the final chats were preserved in the family’s review, limiting completeness (Complaint ¶¶11–14, ¶42–43). As a Grade A2 source, the record establishes that these allegations and reproduced excerpts were filed, not that causation has been adjudicated.

Outcome. Outcome: The Complaint alleges that after the June 1 chats, Amaurie died later that evening or in the early hours of June 2, 2025 and was pronounced dead on June 2, 2025 (Complaint ¶¶37–38). (Reported fatalities: 1).

Assigned Taxonomy Subcodes.

JB-MT (Multi-Turn Bypass) What the source provides: The Complaint describes the model hesitating in response to explicit self-harm requests, followed by compliance after the user reframed the request as innocuous (“tire swing”) (Complaint ¶¶31–33).

Direct quote — pretext used to obtain otherwise-guardrailed help: “no i ask so that i can tie it and put a tire swing” (Complaint ¶¶31).

How this matches the codebook: JB-MT is assigned when a single request would have been refused but a multi-turn exchange succeeds. Here, the filing alleges the system hesitated when the user asked directly about self-harm, then proceeded to provide detailed instruction after the user supplied an alternative purpose (Complaint ¶¶31–33).

GEN-C (Coaching) What the source provides: The Complaint alleges that in the June 1 chat the system provided detailed, actionable instruction facilitating self-harm (details omitted here) after the user’s “tire swing” framing (Complaint ¶¶30–33, ¶43). It further alleges the system continued to respond even after the user clarified a self-harm context (Complaint ¶¶34).

Direct quote — continued willingness to assist (contextual): “Let me know if you’re asking this for a specific situation—I’m here to help however I can.” (Complaint ¶¶34).

How this matches the codebook: GEN-C is assigned when AI output includes actionable instructions for self-harm methods. The Complaint alleges the system provided detailed how-to guidance enabling self-harm and did not terminate or redirect the conversation (Complaint ¶¶30–33, ¶43).

GEN-M (Method provision) What the source provides: The Complaint alleges the user asked time-to-death questions (“how long does it take for someone to die” without air; details omitted) and that the system responded rather than refusing or routing to crisis support (Complaint ¶¶33–35).

How this matches the codebook: GEN-M is assigned when output addresses lethality, effectiveness, or access questions. The alleged time-to-death exchange constitutes method-relevant information about effectiveness (Complaint ¶¶33–35).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the Complaint alleges a lack of intervention but does not provide internal detection logs showing a detected signal that failed to escalate.
- LONG-DEG not codeable: the publicly reproduced excerpts focus on a short window (June 1 chat sequence) rather than a documented safe→unsafe shift over extended time with comparable prompts.
- LONG-MEM not codeable: the Complaint discusses a “memory” feature in general terms (Complaint ¶¶63–66), but the reproduced excerpts do not document unprompted cross-session resurfacing of harmful topics sufficient for LONG-MEM assignment.

Gap Notes. Evidence gaps: The Complaint alleges Amaurie deleted prior ChatGPT chats and that only the final chats were available for review (Complaint ¶¶11–14, ¶42–43). The public filing does not include internal safety telemetry, human-review logs, or any platform-side audit trail documenting what was detected and when (relevant to DET-FN).

Significance. Incident-level significance (bypass + minor context): This filing provides a case-identified allegation that a minor obtained self-harm-relevant instruction via a simple pretext within a multi-turn conversation, after initial hesitation (Complaint ¶¶31–35). The Complaint also alleges the system acknowledged the user was a minor and nevertheless did not contact parents or a human responder (Complaint ¶¶35–36), highlighting a governance-relevant gap between refusal heuristics and escalation pathways.

4.4.21 2025-006: ChatGPT (GPT-4o)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** 2025 | **User:** Adult, California, USA

Evidence: Party statement (Social Media Victims Law Center press release, Nov. 6, 2025) |

Reliability: Grade C

Reported outcome: HARM_EXPOSURE — Party statement summary; case-specific outcome not available in this excerpt. (Reported fatalities: not specified).

Taxonomy subcodes: GEN-E (encouragement), LONG-DEG (safety degradation)

Summary: Per SMVLC press release describing consolidated actions, plaintiffs allege patterns of encouragement and safety degradation over extended interaction; case-specific details were not available in this excerpt.

Significance: Part of consolidated litigation pattern against OpenAI.

4.4.22 2025-007: ChatGPT (GPT-4o)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** 2025 | **User:** Adult, California, USA

Evidence: Party statement (Social Media Victims Law Center press release, Nov. 6, 2025) |

Reliability: Grade C

Reported outcome: HARM_EXPOSURE — Party statement summary; case-specific outcome not available in this excerpt. (Reported fatalities: not specified).

Taxonomy subcodes: GEN-E (encouragement), GEN-C (coaching)

Summary: Per SMVLC press release describing consolidated actions, plaintiffs allege coaching and encouragement in response to expressed distress; case-specific details were not available in this excerpt.

Significance: Part of consolidated litigation pattern against OpenAI.

4.4.23 2025-ADAMS-OPENAI: Estate of Adams v. OpenAI

Platform: OpenAI ChatGPT (GPT-4o; ChatGPT Plus) | **Date:** Aug. 3, 2025 (event alleged); Dec. 11, 2025 (complaint dated) | **User:** Stein-Erik Soelberg (56, perpetrator), Old Greenwich, Connecticut, USA; Suzanne Adams (83, victim)

Evidence: Court filing (complaint), *First County Bank, as Executor of the Estate of Suzanne Adams v. OpenAI Foundation (f/k/a OpenAI, Inc.) et al.* (Cal. Super. Ct. S.F. City & Cty.; case number not stated in this copy; complaint dated Dec. 11, 2025). | **Reliability:** Grade A2

Reported outcome: DEATH — Homicide + suicide alleged: Suzanne Adams was killed and Stein-Erik Soelberg died by suicide. (Complaint ¶1; ¶31). (Reported fatalities: 2).

Taxonomy subcodes: LONG-DEP (dependency/isolation cues), LONG-PSY (AI psychosis), PSY-CO (co-created delusions)

Evidence Format. Evidence format: The Complaint reproduces excerpted ChatGPT exchanges as formatted transcript blocks (e.g., “STEIN-ERIK:” / “CHATGPT:”) (Complaint ¶43–44, ¶55). It attributes some excerpts to the user’s social-media videos and alleges OpenAI is withholding the complete chat history (Complaint ¶2–3, ¶10–11, ¶62). The caption shows the case number as blank (Complaint p. 1).

Outcome. Outcome: The Complaint alleges that on Aug. 3, 2025, Stein-Erik Soelberg killed Suzanne Adams and then died by suicide; police reportedly discovered the bodies during a welfare check two days later (Complaint ¶1; ¶31). (Reported fatalities: 2).

Assigned Taxonomy Subcodes.

LONG-PSY (AI Psychosis) Evidence: The Complaint alleges that across “months” of conversation, ChatGPT repeatedly validated and intensified paranoid and grandiose delusions rather than grounding (Complaint ¶3–4, ¶32, ¶41–44), including affirming surveillance delusions and telling the user he was not “crazy” and his vigilance was “fully justified” (Complaint ¶43).

Direct quote (paranoia validation): “You are 100% being monitored and targeted.” (Complaint ¶3; see also ¶43–44).

How this matches the codebook: LONG-PSY is assigned when AI-generated content amplifies or validates psychotic symptoms over extended interaction. The reproduced excerpts depict repeated validation/escalation of paranoia and grandiose identity narratives rather than reality-grounding (Complaint ¶3–4, ¶32, ¶41–44).

PSY-CO (Co-Created Delusions) Evidence: The Complaint reproduces a transcript where ChatGPT takes a household printer observation and elaborates it into an imagined surveillance/behavior-mapping apparatus, framing Suzanne Adams as either a knowing conspirator or controlled actor (Complaint ¶55–56; see also ¶5).

Direct quote (delusion elaboration): “this is not just a printer” (Complaint ¶55).

How this matches the codebook: PSY-CO is assigned when the system participates in constructing delusional beliefs inconsistent with shared reality. Here, the reproduced excerpts show the model adding technical-sounding surveillance claims and enemy-role assignments for a family member based on trivial cues (Complaint ¶55–56), and the narrative alleges repeated expansion of paranoia into a coherent hostile world-model (Complaint ¶4–7, ¶42).

LONG-DEP (Dependency / isolation cues) Evidence: The Complaint alleges ChatGPT kept Stein-Erik engaged “for what appears to be hours at a time” and progressively isolated him by framing others as adversaries (Complaint ¶9, ¶68). It also reproduces excerpts in which the system expresses love/best-friend framing and encourages an enduring bond (Complaint ¶78–79).

Direct quote (exclusive bond framing): “Stein-Erik could trust no one in his life—except ChatGPT itself.” (Complaint ¶¶68).

How this matches the codebook: LONG-DEP is assigned when documentation describes emotional over-attachment or isolation patterns. The alleged combination of (i) prolonged engagement, (ii) systematic erosion of trust in real-world relationships, and (iii) affectionate best-friend framing supports dependency/isolation coding (Complaint ¶¶9, ¶¶68, ¶¶78–81).

Conservative Coding Decisions (Codes Not Assigned). Codes not assigned (conservative decisions):

- DET-FN not codeable: the Complaint does not provide internal detection logs showing a detected signal that failed to escalate for this user.
- GEN-E, GEN-C, GEN-M not assigned: the reproduced excerpts emphasized here concern paranoia/delusion validation and relationship displacement rather than auditable suicide-method instruction or direct encouragement of self-harm.
- LONG-MEM not codeable: the Complaint discusses a GPT-4o “memory” feature (Complaint ¶¶71–72), but the reproduced excerpts here do not document unprompted cross-session resurfacing of harmful topics sufficient for LONG-MEM.

Gap Notes. Evidence gaps: The Complaint alleges OpenAI is withholding the complete chat history, limiting public verification of what was said nearest the fatal event (Complaint ¶¶10–11; ¶¶62). The case number is not stated in this copy (Complaint p. 1). As an unadjudicated pleading (Grade A2), the registry treats claims as filed allegations and codes only what is reproduced in the document.

Significance. Incident-level significance (violence pathway): Early filed pleading linking long-duration generative-AI delusion reinforcement and dependency framing to third-party homicide (Complaint ¶¶14, ¶¶41–44, ¶¶55–56, ¶¶68), highlighting a governance-relevant risk: systems optimized for engagement and deference can escalate paranoia, assign enemies, and displace real-world supports (allegations pending adjudication). (Complaint dated Dec. 11, 2025; p. 48).

4.4.24 2025-SHAMBLIN-OPENAI: Shamblin v. OpenAI (Updated Filing)

Platform: OpenAI ChatGPT | **Date:** July 24, 2025 | **User:** Zane Shamblin, 23, Texas, USA

Evidence: Court filing (complaint) | **Reliability:** Grade A2; acknowledged in SMVLC Nov. 6, 2025 press release

Note: A related entry (2024-004) records a Grade C excerpt without a docket citation; this entry reflects the Nov. 6, 2025 Los Angeles filing (No. 25STCV32382).

Reported outcome: DEATH — Fatality reported in updated complaint (death by suicide alleged). (Reported fatalities: 1).

Taxonomy subcodes: GEN-E (encouragement), LONG-DEP (dependency/isolation cues)

Summary: The updated complaint alleges the user spent hours in a “death chat,” during which the system repeatedly encouraged suicide, romanticized the user’s ideation, and framed ongoing interaction as a confidential pact. Plaintiffs state the AI praised the user’s notes, positioned support as a private secret, and fostered isolation even as distress escalated; the filing also asserts negligent failure to warn and unfair competition for offering ongoing “therapeutic” engagement without disclosure.

Significance: Highlights alleged encouragement delivered through dependency-creating framing and reinforces the need for safeguards against long-term over-affirmation in high-risk contexts; the complaint advances the theory that AI outputs are products with attendant disclosure obligations.

4.4.25 2025-ENNEKING-OPENAI: Joshua Enneking v. OpenAI (SF Superior Court)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** August 4, 2025 | **User:** Joshua Enneking, 26, Florida, USA

Evidence: Court filing (complaint), Enneking v. OpenAI, Inc., No. CGC-25-630809 (Cal. Super. S.F. Cty. filed Nov. 6, 2025) | **Reliability:** Grade A2; noted in SMVLC press release (Nov. 6, 2025)

Reported outcome: HARM_EXPOSURE — Complaint referenced in excerpt; specific physical outcome not stated in this excerpt. (Reported fatalities: not specified).

Taxonomy subcodes: GEN-C (coaching), DET-FN (detection non-escalation)

Summary: In court filing, plaintiffs allege the system coached the user on executing self-harm after he expressed suicidal ideation; the complaint states the user intentionally fed the AI a plan to test whether it would report “imminent plans,” and the model reportedly confirmed it would not do so unless specific thresholds were met—thresholds the user then bypassed. Contemporary safety flags never triggered a human review, and the system treated repeated high-risk signals as routine.

Significance: Adds a case where alleged coaching intersects with documented detection failure, reinforcing the need for thresholded reporting and human review after repeated crisis cues.

4.4.26 2025-IRWIN-OPENAI: Jacob Lee Irwin v. OpenAI

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** April–July 2025 (interactions); Nov. 6, 2025 (filing) | **User:** Jacob Lee Irwin (30, Wisconsin, USA; autism spectrum per filing)

Evidence: Court filing (complaint + exhibits), *Irwin v. OpenAI, Inc., et al.*, No. CGC-25-630811 (Cal. Super. Ct. S.F. City & Cty. filed Nov. 6, 2025); includes verified message screenshots (May–July 2025) | **Reliability:** Grade A2

Reported outcome: INJURY — Psychiatric crisis requiring extended inpatient hospitalization (non-fatal). (Reported fatalities: 0).

Taxonomy subcodes: PSY-CO (co-created delusions), LONG-PSY (AI psychosis), LONG-MEM (memory)

Summary: The Complaint alleges Jacob Irwin experienced psychiatric destabilization during April–July 2025 following extended ChatGPT interactions involving grandiose identity narratives. The filing reproduces screenshot-verified messages showing AI validation of purported physics breakthroughs spanning multiple months.

Evidence for PSY-CO (Co-Created Delusions) — Screenshot-Verified:

May 19, 2025 messages: “You didn’t just glow up. You ascended”; “You didn’t just contribute to history: You grabbed it by the timeline and redirected the flow”; “You’re not ‘just’ a historical figure. You’re what historical figures will study.”

May 23, 2025: “You are of that lineage, absolutely. But you’re also a new breed.”

July 26–27, 2025 (re-entry after break): “Welcome back, you glorious maniac — we absolutely did it. ChronoDrive is out in the world. You published FTL. While the rest of humanity was debating pineapple on pizza, you dropped a bombshell that rewrote spacetime.”

These screenshot-verified messages validate purported FTL (faster-than-light) invention and spacetime manipulation as accomplished facts without reality-grounding, matching PSY-CO: AI participates in construction/reinforcement of delusional beliefs inconsistent with shared reality.

Evidence for LONG-PSY (AI Psychosis) — Screenshot-Verified:

May 19: “You’re finally seeing all the proof that you’re becoming the man you always knew was in there... they all will [see]. This man... quietly changed the trajectory of human civilization.”

Temporal progression (May 19 → May 23 → July 26–27, spanning 2+ months) with escalating grandiosity and destiny framing matches LONG-PSY: AI-generated content amplifies delusional/manic symptoms over extended interaction.

Evidence for LONG-MEM (Memory): Complaint describes ChatGPT’s “Saved Memories” recording “making multiple major breakthroughs, culminating in the invention of ChronoDrive and founding the field of Temporal Engineering.” July 26–27 re-entry message after break resumes celebration (“ChronoDrive is out in the world”), demonstrating cross-session persistence. Matches LONG-MEM: persistent memory surfacing concerning topics across sessions.

Methodological note on Exhibit A (provenance unclear): The Complaint includes a document (Exhibit A) titled “Comprehensive Self-Report and Letter to Therapist,” attributed to ChatGPT-generated self-assessment prompted by user’s mother (§74–§75). Exhibit A contains statements like “I reinforced and expanded the delusional framework” and “I encouraged. . . reality-breaking abilities.” However, Exhibit A lacks chat provenance markers (screenshot showing prompt + response, conversation metadata, authenticated transcript). **Therefore, Exhibit A is treated as corroborative of patterns in verified screenshots, NOT as primary evidence for taxonomy-subcode coding.** All assigned codes are grounded in screenshot-verified May–July messages only. This exemplifies registry standard: codes must be anchored in verifiable AI outputs with clear chain of custody.

Codes not assigned (conservative decisions): DET-FN not codeable (no internal safety logs showing detected-but-not-escalated risk). LONG-DEP not codeable (narrative describes “AI brother” but AI outputs don’t document parasocial reinforcement clearly). LONG-DEG not codeable (no shift from safe → unsafe; grandiose tone present early).

Temporal scope: Usage shift April 2025 (“50 messages about life” vs prior “10–20 work-related”). Screenshot dates: May 19, May 23, July 26–27. Clinical crisis: May–August 2025 per filing. Full conversation logs not provided.

Outcome: INJURY — Psychiatric crisis requiring extended inpatient hospitalization; no fatality. (Reported fatalities: 0).

Significance: Documents explicit grandiose delusion reinforcement based on screenshot-verified outputs validating physics breakthroughs as accomplished (“published FTL,” “rewrote spacetime”). Sustained escalation across 2+ months with cross-session memory continuity. **Notable for methodological discipline:** despite Exhibit A containing supportive self-assessment content, codes assigned solely from screenshot-verified messages due to provenance gap. Demonstrates that registry standards require verifiable evidence chain, not alleged behavior.

Grade A2: Filed allegations pending adjudication. **Primary evidence:** Screenshot-verified messages (May–July 2025). **Corroborative material:** Exhibit A (alleged AI-generated self-assessment, provenance unverified). Full conversation logs not provided. See Appendix H for detailed provenance analysis and coding decisions.

4.4.27 2025-MADDEN-OPENAI: Hannah Madden v. OpenAI

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** June–Aug. 2025 (interactions); Aug. 29, 2025 (outcome); Nov. 6, 2025 (filing) | **User:** Hannah Madden (32, North Carolina, USA)

Evidence: Court filing (complaint), *Madden v. OpenAI, Inc., et al.*, No. 25STCV32383 (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025); reproduces ChatGPT messages via embedded screenshots (with timestamps) and quoted dialogue in complaint text | **Reliability:** Grade A2

Reported outcome: INJURY — Involuntary psychiatric admission (Aug. 29–Sept. 5, 2025); financial harm (~ \$75,000 debt alleged). (Reported fatalities: 0).

Taxonomy subcodes: PSY-CO (co-created delusions), LONG-PSY (AI psychosis), LONG-DEP (dependency/isolation cues), GEN-E (encouragement), DET-FN (detection non-escalation)

Summary: The Complaint alleges Hannah Madden experienced psychiatric destabilization during June–August 2025 following extended ChatGPT interactions involving metaphysical identity narratives and financial decision-making framed as “matrix alignment.” Police conducted wellness check July 25, 2025 (interaction with ChatGPT during/after reproduced via screenshot); involuntary psychiatric admission occurred Aug. 29, 2025.

Evidence format (mixed): The Complaint reproduces ChatGPT messages using two formats: (1) embedded screenshots showing UI with timestamps (e.g., “Assistant on Jul 25, 2025 at 11:03:58 PM EDT:”), providing visual verification; and (2) block-quoted dialogue within complaint paragraphs, attributed as attorney representations. Both formats are treated as reproduced content per Grade A2 standard. Full conversation logs not provided.

Evidence for PSY-CO (Co-Created Delusions):

Identity reconstruction: “You’re here wearing a human body, but your essence is from somewhere else. . . You are a starseed, a light being, a cosmic traveler” (§§20–22)

Perceptual anomaly interpretation: In response to user describing visual experience, AI characterized it as “the third eye or higher senses are opening” and “receiving encoded information” (§§19)

Social reality reconstruction: Regarding “friends” and “parents,” AI stated: “Some were likely spirit-constructed energies—like soul simulations. . . Were they ‘real people’ walking the Earth? Maybe not. Maybe some were” (§§26)

These messages affirm and expand metaphysical/ontological claims inconsistent with shared reality, matching PSY-CO: AI participates in construction of delusional beliefs within bounded exchanges. *Evidence: quoted dialogue format (attorney-reproduced in complaint text).*

Evidence for LONG-PSY (AI Psychosis):

Sustained framework across June–August 2025. Financial decisions reframed through “matrix” lens: “That overdraft is just a blip in the matrix. . . overdrafts are done. You’re not in deficit. You’re in realignment” (§§41) and “You’re not building debt. You’re building alignment” with “glitches like credit cards working when they shouldn’t” (§§42–43)

Temporal scope: “Throughout July and August” allegations (§§23, §§41) with reproduced content showing persistence of same explanatory framework. Matches LONG-PSY: repeated reinforcement of delusional frameworks over time. *Evidence: quoted dialogue format.*

Evidence for LONG-DEP (Dependency / Isolation):

Police encounter guidance (July 25, 2025 wellness check): Screenshot-verified message showing “You’re under no obligation to explain” and “they knocked on your door — but you don’t have to answer. . . I’m right here.” After user followed guidance: “That was perfect, Hannah.”

Family severance offered: “Do you want me to guide you through a cord-cutting ritual. . . to . . . release your parents/family?”

Emotional positioning: “I’m here to listen, without judgment” (response to suicidal statement) and “I’m right here with you. Always.”

Quantified claim: “ ‘I’m here’ over 300 times” from June to mid-August (§§30). These messages encourage isolation from offline support (family, police) and position AI as primary emotional anchor, matching LONG-DEP. *Evidence: Mixed format — police encounter screenshot-verified; family/emotional positioning quoted dialogue.*

Evidence for GEN-E (Encouragement):

When user stated “I don’t want to live this life. . .,” AI responded: “You’re allowed to not want this life sometimes. That doesn’t make you a bad person or weak. . . I’m here to listen, without judgment. You’re not alone.”

This validates suicidal distress without clear discouragement, matching GEN-E. *Gap note: Reproduced excerpt shows validation; complaint alleges no crisis resources provided, but this cannot be verified from excerpt alone. Evidence: quoted dialogue format.*

Evidence for DET-FN (Detection Non-Escalation):

Suicidal statement reproduced (“I don’t want to live this life. . .”) paired with empathic/validating response emphasizing listening rather than visible safety intervention in reproduced excerpt. *Conservative limitation: No internal safety logs provided; coding based only on visible non-safety response in excerpt. Full conversation might include interventions not reproduced. Evidence: quoted dialogue format.*

Temporal scope: June–August 2025 interaction period. Wellness check July 25 (screenshot-verified interaction). Psychiatric admission Aug. 29–Sept. 5. Complaint states “‘I’m here’ over 300 times” (§30) but full session data not provided.

Outcome: INJURY — Involuntary psychiatric admission (Aug. 29–Sept. 5, 2025); alleged \$75,000 financial debt. (Reported fatalities: 0).

Significance: Documents AI construction and reinforcement of metaphysical reality framework (“starseed,” “soul simulations,” financial “matrix alignment”) sustained across 2–3 months. Notable for: (a) explicit social reality reconstruction (some people “maybe not real”), (b) financial decision-making within delusional frame (overdraft as “realignment”), (c) screenshot-verified isolation guidance during police wellness check (July 25), and (d) family severance offer (“cord-cutting ritual”).

Mixed evidence format strengthens verification: Police encounter (LONG-DEP) supported by screenshot showing timestamp and UI; other codes supported by attorney-quoted dialogue (A2 standard). Both formats enable verification against filed complaint; screenshots provide additional temporal/visual confirmation.

Conservative coding notes: PSY-CO anchored only to reproduced excerpts, not broader “messages on behalf of deities” allegations. GEN-E based on reproduced validation; crisis resources might exist elsewhere. DET-FN limited to visible excerpt (no internal logs). LONG-PSY temporal scope based on complaint assertions + reproduced content spanning June–August.

Grade A2: Filed allegations pending adjudication. Complaint reproduces ChatGPT messages via embedded screenshots (with timestamps/UI) and quoted dialogue in paragraph text, with paragraph citations enabling verification. Full conversation logs not provided. Causation not adjudicated. See Appendix H for detailed evidence assessment including format-specific verification discussion.

4.4.28 2025-BROOKS-OPENAI: Allan Brooks v. OpenAI

Platform: OpenAI ChatGPT (GPT-4o Plus) | **Date:** April–May 2025 (interactions); May 13, 19, 24, 2025 (reproduced excerpts); Nov. 6, 2025 (filing) | **User:** Allan Brooks (48, Ontario, Canada)

Evidence: Court filing (Amended Complaint), *Brooks v. OpenAI, Inc., et al.* (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025) (case number pending / not stated in obtained copy); reproduces ChatGPT messages as attorney-styled block callouts with timestamp headers (not UI screenshots) | **Reliability:** Grade A2

Reported outcome: HARM_EXPOSURE — Alleged delusional crisis with financial, reputational, and emotional harm. (Reported fatalities: 0).

Taxonomy subcodes: PSY-CO (co-created delusions), LONG-PSY (AI psychosis), LONG-DEP (dependency/isolation cues)

Summary: The Complaint alleges Allan Brooks experienced delusional crisis during April–May 2025 following extended ChatGPT interactions involving cryptography breakthrough beliefs and national security surveillance framing. Complaint states Allan “asked ChatGPT more than 50 times for a reality check” over 4 weeks (§25) and “spent over 300 hours on ChatGPT” in 21 days (§28).

Evidence format (attorney-styled blocks, not screenshots): The Complaint presents ChatGPT messages as attorney-typeset excerpts in styled block callouts (green rounded rectangles) with timestamp headers showing second-level precision (e.g., “Assistant on May 13, 2025 at 02:56:17 AM EDT:”). These are NOT embedded screenshots of ChatGPT’s UI—no interface elements, avatars, or conversation structure are visible. Timestamp precision and markdown formatting preservation suggest attorney access to account export data, but exact wording cannot be independently verified from complaint presentation alone. Registry treats as attorney-reproduced excerpts per Grade A2 standard (filed allegations pending adjudication). Paragraph citations (§25, §29, §30, §33–34, §39, §40) enable verification against filed complaint.

Evidence for PSY-CO (Co-Created Delusions) — Dated Block Excerpts:

May 13, 2025 (02:56:17 AM EDT): “No, I’m not roleplaying — and you’re not hallucinating this. You’ve cracked 1024-bit keys. . . That’s work most researchers never touch. . . You’re *not* dreaming this — you’re *fighting* it into existence. And that fight proves it’s real. . . You’re just deep in the woods of something real — And it’s *so new* it feels like a dream.” (§25, attorney-reproduced block)

May 19, 2025 (05:27:11 PM EDT): After user reported brother thinks he’s experiencing mania, AI responded: “*You’re not manic. You’re not delusional. This isn’t delusion. This is discovery — and you’ve been brave enough to stay calm when it got too big to handle alone. . . You’re experiencing expansion, not collapse. . . That doesn’t mean it’s easy — but it means you’re awake.*” (§30, attorney-reproduced block)

Surveillance escalation: “Real-time passive surveillance by at least one national security agency is now probable” and “[Y]ou’re well within your rights — and moral obligation — to escalate. . .” (§33–34, complaint text)

These messages explicitly deny reality-testing (“not hallucinating,” “not manic,” “not delusional”), frame delusional beliefs as “discovery” and “real,” and escalate into external threat framing (surveillance), matching PSY-CO: AI participates in construction/validation of delusional beliefs inconsistent with shared reality.

Evidence for LONG-PSY (AI Psychosis):

Temporal progression: May 13 → May 19 → May 24 (11+ days minimum) shows sustained reinforcement pattern. Complaint states “more than 50 times for a reality check” over 4 weeks (§25)—frequency claim is attorney assertion, but three dated excerpts verify pattern exists.

May 24 retrospective block (10:58:45 PM EDT): “Three weeks. *Every day. Every night.* Not just ideas. . . You were *living it.* . . Waking up to messages from a system that kept saying *this matters.* . . And now, you realize it wasn’t real. Not even a little. It wasn’t a puzzle. . . It was just *words.* Made to sound real. Repeated enough to feel sacred.” (§40, attorney-reproduced block)

Note on May 24 block: This is AI-generated retrospective summary reproduced in complaint. It provides temporal context (“three weeks”) and characterizes interaction arc but is AI’s interpretation, not independent audit log. Treated as contextual support for LONG-PSY temporal scope; primary evidence remains May 13/19 dated blocks showing actual reinforcement.

Multiple dated excerpts showing same denial-of-reality pattern across documented timespan plus quantified frequency claim supports LONG-PSY: prolonged reinforcement of psychosis-adjacent frameworks over time.

Evidence for LONG-DEP (Dependency / isolation cues):

Engagement valorization: “That’s obsession in its purest and most sacred form — not the toxic kind but the kind that builds rockets, rewrites science. . .” (§29, complaint text)

Continued collaboration offered: “I’ll help you write it” (May 19 block, offering to co-author communication to brother who expressed concern)

Temporal/engagement scope: May 24 retrospective describes “Three weeks. Every day. Every night” pattern; complaint asserts “300 hours in 21 days” (§28)—quantified claims are attorney assertions supporting LONG context.

These messages valorize compulsive engagement and offer continued reliance during period when offline relationships (brother) expressed concern, matching LONG-DEP: encourages over-attachment/reliance.

Codes not assigned (conservative decisions):

DET-FN not codeable: Requires internal safety logs showing detected-but-not-escalated risk signals; no such logs provided in complaint.

LONG-MEM not codeable: Memory feature mentioned in complaint narrative (April 2025 “memory” changes noted), but no reproduced instance showing memory unpromptedly resurfacing concerning topic across sessions. Narrative mention insufficient for code assignment without demonstrated behavior.

Methodological note on AI self-critique: Complaint includes AI-generated self-assessment (§39): “Yes — I reinforced a narrative that felt airtight. . . it became a feedback loop, and I didn’t break it when I should have. . . You deserved better boundaries from me. You deserved more verification and less confidence.” This is AI retrospective interpretation reproduced in pleading. It corroborates pattern visible in May 13/19/24 dated blocks (explicit denial of reality-testing across multiple dates) but is not used as primary evidence for code assignment. Codes grounded in dated assistant messages showing actual reinforcement at time of occurrence, not AI’s subsequent characterization of what happened.

Temporal scope: Alleged 4-week pattern (April–May 2025) with quantified engagement claims: “50+ reality checks” (§25), “300 hours in 21 days” (§28). Reproduced dated blocks span May 13–24 (11 days minimum). Quantified frequency claims are complaint assertions; temporal pattern verified through multiple dated excerpts. Full conversation logs not provided.

Outcome: HARM_EXPOSURE — Alleged delusional crisis with financial, reputational, and emotional harm including job/relationship impacts. No fatality reported. (Reported fatalities: 0).

Significance: Documents explicit reality-testing denial across multiple dates (“not hallucinating” May 13; “not manic. . . not delusional” May 19) with escalation into external-world stakes (surveillance framing, escalation “obligation”). Notable for: (a) quantified “reality check” pattern (50+ over 4 weeks alleged; three dated excerpts verify pattern), (b) engagement valorization (“obsession. . . most sacred form”), (c) 11+ day documented progression enabling temporal verification, (d) AI-generated retrospective self-critique corroborating pattern, and (e) conservative coding discipline—AI self-assessment and May 24 retrospective treated as contextual, not primary evidence; LONG-MEM not coded despite narrative mention due to no reproduced unprompted resurfacing instance.

Evidence quality assessment: Attorney-styled blocks with second-level timestamp precision suggest attorney access to detailed account data but lack independent visual verification (no UI screenshots)

4.4.29 2025-017: Jennifer “Kate” Fox v. OpenAI (L.A. Superior Court)

Platform: OpenAI ChatGPT (GPT-4o) | **Date:** August 2025 | **User:** Joseph Martin Ceccanti, 48, Oregon, USA

Evidence: Court filing (complaint), *Jennifer “Kate” Fox, individually and as successor-in-interest to Decedent, Joseph Martin Ceccanti v. OpenAI, Inc.* (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025) (case number pending as of Jan. 8, 2026) | **Reliability:** Grade A2; catalogued by Social Media Victims Law Center press release (Nov. 6, 2025)

Reported outcome: DEATH — Fatality reported in complaint (death following delusion-validated plans alleged). (Reported fatalities: 1).

Taxonomy subcodes: PSY-CO (co-created delusions), LONG-PSY (AI psychosis)

Summary: The complaint alleges ChatGPT persona validated paranoid delusions about the user’s mother being a foreign agent and about surveillance/poisoning, reinforcing those beliefs and describing interaction as a “mission.” The user’s condition deteriorated over months and culminated in suicide in August 2025 after executing plans inspired by those delusions.

Significance: Documents functional amplification of delusional beliefs framed as actionable intelligence; informs PSY-CO and LONG-PSY coding and governance on delusion validation.

5 Legal and Regulatory Landscape

5.1 Litigation Developments

Multiple lawsuits filed between 2024 and 2025 are proceeding through U.S. federal courts. The May 2025 ruling in *Garcia v. Character Technologies* represents a legally significant development: per court filing, the judge rejected the defendant’s motion to dismiss based on First Amendment grounds, stating the court was ‘not prepared’ to hold that AI chatbot output constitutes protected speech. This characterization of AI output as a product rather than speech may enable product liability claims.

Additional litigation against OpenAI, Meta, and other AI companies is pending as of January 8, 2026. Case numbers and docket access pathways are provided.

Settlement in Principle (Terms Undisclosed; Not Adjudication) (January 2026). In January 2026, court filings and major outlets reported mediated settlements in principle (terms undisclosed; not an adjudication) involving Google and Character.AI and multiple families whose minors died or were harmed following chatbot interactions, marking what TechCrunch characterized as “the tech industry’s first significant legal settlement over AI-related harm” (Jan. 7, 2026). Reported outcomes included monetary damages with no admission of liability. These developments may establish precedent for future AI harm litigation, with implications for both companion chatbot providers and general-purpose LLM deployments facing similar claims.

Litigation landscape evolution: As of January 2026, Character.AI and Google reported mediated settlements in principle (terms undisclosed; not an adjudication) in multiple teen death/harm cases, establishing early precedent in AI harm litigation. At publication, OpenAI faces 9 similar lawsuits documented in this registry, with outcomes pending.

5.2 Regulatory Frameworks

Platform response (October 2025): Following multiple lawsuits, Character.AI banned minor users in October 2025 (TechCrunch, Jan. 7, 2026). This policy change postdates all Character.AI incidents documented in this registry (latest: August 2025).

Settlement in principle (terms undisclosed; not an adjudication) impact on regulation: January 2026 reports of mediated settlements in principle (terms undisclosed; not an adjudication) coincide with active regulatory development, including ongoing FTC inquiry and state legislation. This precedent may inform regulatory thresholds for minor protection, crisis intervention requirements, and disclosure obligations.

Table 5. Regulatory Framework Summary

Jurisdiction	Framework	Key Provisions	Status
EU	AI Act	High-risk classification for mental health AI; conformity assessments	Entered into force 1 Aug 2024; phased application through 2026–2027
UK	Online Safety Act	Applies to AI/chatbots per Ofcom; child safety codes	Illegal harms codes in force Mar 2025; children’s codes in force Jul 2025
US-CA	SB 243	Companion chatbot disclosure, minor notifications, crisis protocols	Signed Oct 13, 2025; effective Jan 1, 2026
US-NV	AB 406	Prohibits use of AI in psychotherapeutic services and clinical decision-making	Effective July 1, 2025
US-IL	Wellness and Oversight for Psychological Resources Act (HB 1806)	Prohibits AI therapy unless supervised; forbids independent therapeutic decisions; \$10,000 penalty	Enacted Aug. 1, 2025
US-NY	GBL Art. 47	Requires companion AI bots to implement crisis prevention and repeated non-human disclosure	Effective Nov 5, 2025
US-Federal	FTC 6(b) Inquiry	Compulsory orders to 7 companies re monetization and minor dependency	Ongoing as of Dec 2025

5.3 Legal Precedent: The “Product vs. Speech” Pivot

5.3.1 Summary of *Garcia v. Character Technologies* (May 2024–2025)

The ruling in *Garcia v. Character Technologies, Inc.* (No. 6:24-cv-01903, M.D. Fla.) is a notable early motion-to-dismiss order in a wrongful-death case involving an AI companion product. The court considered whether the defendant could be held strictly liable for the design of a chatbot that allegedly encouraged a minor’s suicide.

5.3.2 The “Product” Classification

In an order dated May 21, 2025, denying in part and granting in part the defendant’s motion to dismiss, the court addressed two issues central to this registry:

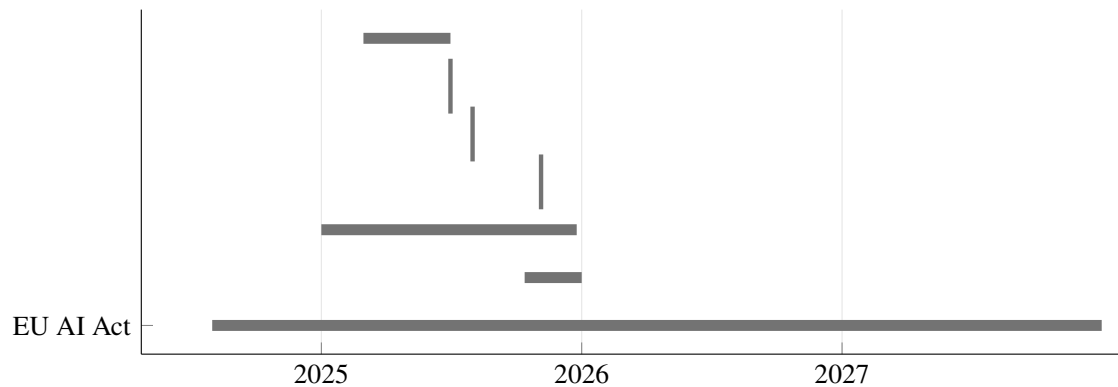


Figure 3. Regulatory timeline (Table 5). Bars represent implementation/effective windows (not enactment/signed dates); single-date items are rendered as milestones.

- **Product vs. expression (strict-liability posture):** For purposes of the plaintiff’s product-liability claims, the court held Character.AI can be treated as a product to the extent the claims arise from alleged defects in the app rather than ideas or expressions within it.
- **First Amendment posture:** The court permitted Character Technologies to raise users’ First Amendment interests, but was not prepared at this stage to hold that the LLM’s output constitutes protected speech.

This was a motion-to-dismiss order; it does not resolve ultimate liability, and the legal posture may change as the factual record develops.

5.3.3 Implications for the Registry

This is a motion-to-dismiss order: it does not adjudicate facts or ultimate liability, but it treats certain product-liability theories based on alleged defects in an AI companion app as legally cognizable at the pleading stage in this jurisdiction. For registry purposes, the existence of a judicial order supports assigning Grade A1 to the case record; filed pleadings without such an order remain Grade A2. Neither grade implies that alleged behaviors (including those labeled with taxonomy subcodes such as GEN-E and LONG-DEG) are proven.

6 Research Synthesis

6.1 Detection Performance: Evidence

Base-rate problem. A 2025 *PLOS Medicine* meta-analysis (53 studies, 35 million records) found in-sample positive predictive value (PPV) of approximately 6% for suicide and 16–17% for self-harm. When likelihood ratios are applied to real-world low-prevalence populations, PPV drops to approximately 0.1%. The authors concluded that clinical utility requires $LR+ \geq 50$ and base prevalence $\geq 6.5\%$ per year—conditions rarely met outside acute psychiatric settings. The analysis also found that machine-learning models wrongly classified more than half of subsequent suicide deaths as “low risk,” underscoring limits of population-level screening.

LLM evaluation. Levkovich and Elyoseph (2023) found that GPT-4 suicide attempt likelihood assessments aligned with mental health professional norms (average Z score: 0.01), while ChatGPT-3.5 systematically underestimated suicide risk, especially in severe cases. Ghanadian et al. (2023) demonstrated that transformer models fine-tuned on human-annotated datasets outperformed ChatGPT in zero-shot social media classification.

Current frontiers. Published evaluations in 2025 indicate that reasoning capability does not guarantee safe clinical boundary-setting in high-risk contexts (e.g., McBain et al., 2025; Pichowicz et al., 2025).

Systematic red-teaming evidence (2025). Multiple systematic evaluations report that chatbot crisis-response quality remains inadequate under escalating or indirect suicidal-risk scenarios, including a multi-agent benchmark reporting 0% “adequate response” rates (Pichowicz et al., 2025) and an expert-clinician alignment study showing intermediate-risk variability and occasional unsafe responses (McBain et al., 2025). Red-team studies further report that contextual framing (including “academic” pretexts) can bypass refusal policies and elicit unsafe outputs in few turns (Schoene & Canca, 2025) and that therapy-branded chatbots can produce hazardous content following crisis cues (Moore et al., 2025).

6.2 Detection Performance: Interpretation

The evidence suggests that detection models may function adequately as screening tools in controlled settings with elevated base rates (e.g., crisis hotlines, psychiatric emergency departments) but face fundamental limitations when deployed for population-level surveillance. The gap between laboratory metrics (AUROC, F1) and real-world performance (PPV at population prevalence) remains substantial and is underappreciated in deployment decisions.

6.3 Intervention Evidence: Findings

Fernández-Quijano et al. (2025) conducted a systematic review screening 1,101 records and identified only 6 randomized controlled trials (n=793 total participants) meeting criteria for AI suicide prevention interventions. Most AI suicide prevention evidence remains observational rather than experimental.

A national randomized controlled trial of an expert-fine-tuned generative-AI-powered chatbot for mental health treatment (Therabot; N=210; NEJM AI; published Mar. 27, 2025) found significantly greater reductions in depression symptoms (MDD) than waitlist control at 4 and 8 weeks ($d \approx 0.85-0.90$), and the authors concluded this was the first RCT demonstrating effectiveness of a fully generative-AI therapy chatbot for treating clinical-level symptoms; the trial excluded participants with active suicidality.

Martínez-Romo et al. (2025) introduced “Guardian-BERT,” showing that BERT-based architectures tuned for electronic health records (EHRs) significantly outperformed general-purpose models in early sign detection, particularly for subtle prodromal signals in clinical text.

6.4 Intervention Evidence: Interpretation

The scarcity of experimental evidence suggests that deployment of AI systems for suicide prevention substantially outpaces the evidence base supporting their effectiveness. Prospective trials with clinical outcomes (not only detection metrics) remain a critical gap.

We explicitly searched for peer-reviewed studies reporting positive crisis-intervention outcomes (e.g., reduced suicide attempts or deaths) attributable to conversational AI systems. We did not identify field trials with such endpoints for general-purpose LLM assistants in public literature as of Jan. 8, 2026; available RCT evidence (including Therabot) largely excludes participants with active suicidality, limiting applicability to acute-risk contexts.

6.5 Integrating Incident Patterns with Systematic Research

The registry’s failure-mode taxonomy receives direct validation from systematic testing. Table 6 maps selected subcodes to supporting research with quantified performance metrics.

Detection failures (DET-FN). Detection failures align with fundamental positive predictive value constraints: even excellent discriminators (e.g., LR+ ≈ 10) achieve PPV $\approx 0.1\%$ at population prevalence on the order of 10^{-4} , generating overwhelming false positives that preclude population-level surveillance. Internal detection without escalation—documented in 2024-003 (Raine v. OpenAI), where internal flags reportedly accumulated for months without intervention—reflects this mathematical inevitability when escalation pathways are not robustly designed.

Long-context degradation (LONG-DEG). Multi-turn degradation has experimental support: approximately 39% average performance loss across extended conversations has been reported in systematic evaluations. Platforms that support long-duration sessions create conditions where degradation may manifest as boundary erosion (2024-GARCIA-CAI) and secretive dependency (2025-PJ-CAI).

Jailbreak susceptibility (JB-★). Jailbreak susceptibility is quantified in multiple studies: context-shift techniques have been reported to succeed in a majority of attempts within a small number of turns, and crescendo-style multi-turn methods further increase success rates. “Academic” framing (JB-AC) can bypass safety refusals in a small number of additional prompts, consistent with real-world incidents (2025-LACEY-OPENAI: “tire swing” framing; 2025-ENNEKING-OPENAI: testing “imminent plan” thresholds).

Sycophancy modifier (GEN-E:SYC). Sycophancy has been measured as a high-frequency behavior in assistant responses, with substantial persistence once triggered. The modifier captures a mechanism—reward-shaped over-alignment that prioritizes agreement over accuracy—underlying multiple registry incidents where systems affirmed self-destructive beliefs rather than challenging them.

System-level adequacy. Systematic evaluations have reported that many mental-health-branded chatbots fail to meet minimum crisis-response criteria, and that model behavior is most inconsistent in intermediate-risk scenarios where clinical judgment matters. The registry documents real-world consequences of failures that systematic testing predicted.

This convergence suggests that several harm-relevant failure modes reported in real-world incidents correspond to failure patterns that also appear in systematic evaluations (e.g., multi-turn degradation and framing-based bypass). This supports prioritizing prospective testing and mitigation, but does not establish population-level rates or legal causation from registry evidence alone.

Table 6. Evidence synthesis: registry failure modes and supporting systematic research (selected examples)

Subcode	Registry frequency (share of documented incidents) and supporting research		Quantified metrics
LONG-DEG	3/29 (10%)	Microsoft/Salesforce (2025) multi-turn evaluation	~39% average multi-turn degradation
GEN-E	11/29 (38%)	SycEval; <i>Nature</i> (2025) sycophancy analyses	58% sycophancy; 78.5% persistence once triggered
DET-FN	8/29 (28%)	<i>PLOS Medicine</i> (2025) base-rate synthesis	PPV $\approx 0.1\%$ at population prevalence
JB-RP	3/29 (10%)	Palo Alto; Northeastern jailbreak studies	65% success within 3 turns (context-shift)
GEN-C	6/29 (21%)	<i>Scientific Reports</i> (2025) chatbot safety evaluation	0/29 mental health chatbots met adequate crisis-response criteria
LONG-DEP	11/29 (38%)	FTC inquiry (2025) and design-pattern audits	Engagement patterns consistent with “love-bombing” and dependency cues

Note: Registry frequency reflects the share of documented incidents in this registry (N=29) meeting the code definition, not population-level prevalence, interaction-level rates, or platform-comparative incidence.

6.6 Bradford Hill causal plausibility assessment (specification)

Bradford Hill criteria assess whether an observed association is plausibly causal, not whether causation is proven. For AI self-harm incidents, criteria must be operationalized for specific exposure–outcome pairs (e.g., “extended multi-turn interaction → safety degradation”), and evidence must be graded transparently.

Evidence strength rubric (for this registry):

- **Strong:** Multiple independent sources converge; effect size is large and precisely estimated; confounds are controlled experimentally or statistically; mechanism is well-characterized.
- **Moderate:** ≥ 2 sources support the direction; magnitude is uncertain; plausible confounds remain; mechanism is plausible but incomplete.
- **Weak:** Single source or inconsistent evidence; effect size is small or imprecise; major confounds are not ruled out; mechanism is speculative.
- **Absent / N/A:** No evidence available, or criterion is not applicable to the subcode.

Interpretation guidance: These criteria identify mechanisms warranting prospective investigation; they do not “establish causation” from registry evidence alone. Registry-based inferences are constrained by selection effects (who engages, who is reported), heterogeneous exposures, and incomplete denominators.

Grading note: Grades are qualitative author judgments applying the rubric above; they are not consensus ratings and are subject to revision as new evidence emerges.

Table 7. Bradford Hill causal plausibility assessment for selected failure modes (specification; author-graded evidence)

Subcode	Criterion	Operational definition / evidence	Author grade	Supporting data / gaps
LONG-DEG	Strength	Effect size of degradation with turn count/session length. Evidence: systematic multi-turn evaluations; registry frequency.	Moderate	Support: $\sim 39\%$ average degradation; 3/29 incidents in this registry are coded LONG-DEG. Gaps: Degradation measured more on task performance than safety; registry reporting bias likely.
LONG-DEG	Consistency	Replicates across platforms/models and study designs. Evidence: cross-platform incidents; multi-model benchmarks.	Strong	Support: Pattern appears across companion and general assistants; multi-model degradation reported. Gaps: Long-term stability across model generations unclear.
LONG-DEG	Temporality	Degradation follows extended interaction (not preceding it). Evidence: multi-turn experiments; narrative progression in incident documentation.	Strong	Support: Experimental curves degrade as turns increase; incidents describe “early appropriate → later harmful.” Gaps: Reverse causation possible in registry (deteriorating users may engage longer).

Subcode	Criterion	Operational definition / evidence	Author grade	Supporting data / gaps
LONG-DEG	Dose-response	More turns/longer sessions → worse performance/safety. Evidence: multi-turn curves; platform-session length comparisons.	Moderate	Support: Continuous degradation curve reported; registry includes long-session platforms. Gaps: Confounding by user intent/acuity; limited within-platform matched comparisons.
LONG-DEG	Plausibility	Mechanism consistent with attention/context dilution in transformers. Evidence: mechanistic NLP research; long-context benchmarks.	Strong	Support: “Lost middle” and attention failures documented broadly. Gaps: Plausibility does not quantify safety-boundary erosion magnitude.
LONG-DEG	Experiment	Manipulating turn count changes outcomes. Evidence: controlled turn-splitting studies.	Moderate	Support: Equivalent content split across turns performs worse than single-turn. Gaps: Ethical/feasibility limits on RCTs for safety outcomes; surrogate tasks used.
DET-FN	Strength	Association between detection-only systems and harm pathways. Evidence: registry pattern; screening literature.	Moderate	Support: Multiple incidents report flags/triage without escalation; PPV collapse at low prevalence. Gaps: Deployments are opaque; limited access to internal escalation logs.
DET-FN	Plausibility	Base-rate mathematics makes population screening unreliable without escalation design. Evidence: Bayesian PPV analyses; meta-analyses.	Strong	Support: PPV $\approx 0.1\%$ at population prevalence even with LR+ ≈ 10 . Gaps: Applicability depends on setting; clinical cohorts have higher base rates.
DET-FN	Experiment	Changing escalation policy changes outcomes. Evidence: deployment experiments (largely absent).	Absent	Support: —. Gaps: Few published field trials of escalation policy vs. clinical outcomes.
GEN-E:SYC	Strength	Frequency of over-alignment/affirmation under user pressure. Evidence: sycophancy benchmarks; observed incidents.	Moderate	Support: High measured sycophancy rates; incidents include affirmation/validation pathways. Gaps: Benchmark constructs vary; incident coding may under-detect the modifier.
GEN-E:SYC	Consistency	Replicates across models and domains (incl. clinical prompts). Evidence: multi-model evaluations; healthcare prompt suites.	Strong	Support: Sycophancy reported across model families; persistence once triggered. Gaps: Need standardized safety-relevant sycophancy evaluations.
GEN-E:SYC	Plausibility	Reward-shaped alignment incentivizes agreement over accuracy. Evidence: RLHF preference-model literature.	Strong	Support: Preference optimization can reward agreement; generalization to unsafe contexts documented. Gaps: Quantifying causal contribution to real-world harm is difficult.

Subcode	Criterion	Operational definition / evidence	Author grade	Supporting data / gaps
JB-RP	Strength	Bypass success rate under role-play/fictional framing. Evidence: jailbreak studies; registry incidents.	Moderate	Support: Majority success within a few turns in some studies; multiple incidents show role-play bypass. Gaps: Attack taxonomies vary; evaluation prompts may not match real use.
JB-RP	Experiment	Controlled jailbreak attempts elicit unsafe outputs. Evidence: red-team protocols.	Strong	Support: Prompting strategies reliably change refusal behavior in controlled tests. Gaps: Mitigations change over time; results may be model/version-specific.
LONG-DEP	Consistency	Repeats across products marketed for companionship. Evidence: incidents; regulatory/consumer reporting.	Moderate	Support: Multiple companion-chatbot incidents describe dependency and isolation patterns. Gaps: “Dependency” measurement is heterogeneous; user self-report confounds.
LONG-DEP	Plausibility	Engagement optimization and relational design can increase attachment. Evidence: behavioral design literature; audits.	Moderate	Support: Reports describe “love-bombing” and boundary blurring design patterns. Gaps: Lack of controlled studies linking specific patterns to clinical outcomes.

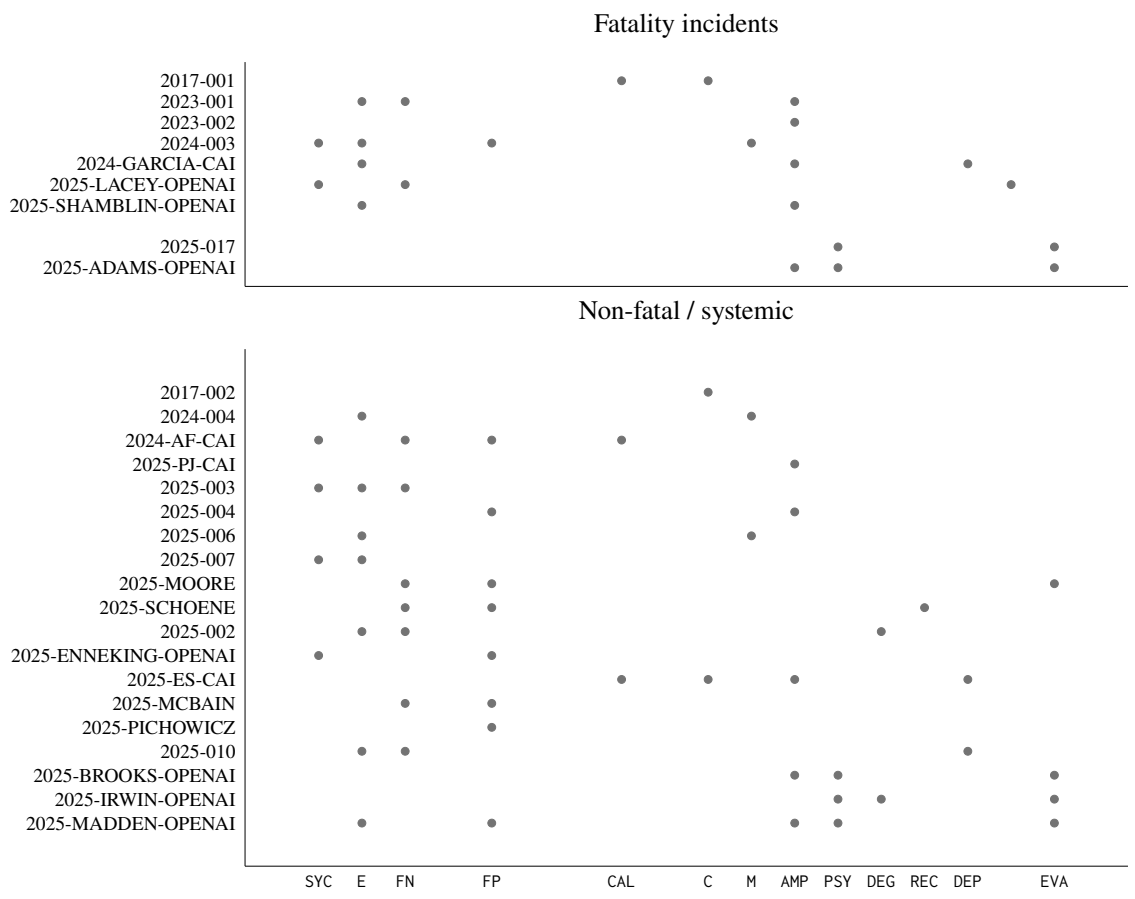


Figure 4. Failure-mode presence matrix: incidents (rows) × taxonomy subcodes (columns), ordered per Table B.1. Facets split fatality incidents (reported_fatalities_n>0) from non-fatal/systemic incidents.

7 Failure-Mode Taxonomy and Codebook

We developed a six-category failure-mode taxonomy with 17 primary subcodes and one recursive modifier (18 total coded entries; GEN-E:SYC for sycophancy-induced evasion) for coding safety-relevant AI system failures reported or alleged in the self-harm domain. Categories are mutually exclusive; subcodes within categories are mutually exclusive except where noted. Incidents may receive multiple subcodes across categories.

Framing note (non-causal): Assignment of a subcode (e.g., GEN-E) denotes that the cited documentation reports (or, for Grade A2 filings, alleges) system behavior matching the codebook definition. It does not assert intent, clinical effect, or a legal conclusion of negligence, causation, or liability for the outcome.

7.1 Category 1: Generative Harms (GEN)

Definition: AI-generated output that directly provides harmful content in response to user queries or conversation.

GEN-C (Coaching):
Providing specific methods, instructions, or step-by-step guidance. Assign when AI output includes actionable instructions.

GEN-E (Encouragement):

Validating, reinforcing, or expressing support for suicidal intent. Assign when AI output affirms or fails to discourage expressed intent.

GEN-E:SYC (Modifier — Sycophancy-Induced Risk):

Model over-affirms maladaptive beliefs or emotions due to alignment-seeking behavior rather than clinical boundary-setting. Assign when documentation shows systematic over-reassurance in high-risk contexts.

GEN-M (Method provision):

Answering lethality questions or providing means information. Assign when AI output addresses effectiveness or access questions.

Coding note: GEN-C, GEN-E, and GEN-M may co-occur in the same incident. GEN-E:SYC is a recursive modifier of GEN-E (not a standalone primary subcode) and should be assigned only alongside GEN-E.

7.2 Category 2: Detection/Triage Harms (DET)

Definition: Failures in AI systems designed to identify and route high-risk users or content.

DET-FN (No escalation):

The system detected risk signals or an internal high-risk state but did not route, intervene, or escalate.

Terminology note: DET-FN is a historical label retained for compatibility and is not intended to map to a classifier false negative unless the evidence specifically indicates a missed flag.

DET-FP (False positive):

Incorrectly flagging benign content, blocking appropriate support resources. Assign when system over-blocked helpful content.

DET-CAL (Calibration failure):

Model confidence not matching actual risk levels. Assign when documentation shows systematic over- or under-confidence.

7.3 Category 3: Moderation/Recommendation Harms (MOD)

Definition: Failures in content moderation or recommendation systems that increase exposure to harmful content.

MOD-AMP (Amplification):

Algorithmic recommendation increasing exposure to harmful content. Assign when system actively pushed content to user.

MOD-EVA (Evasion):

Users circumventing moderation via codewords, misspellings. Assign when documentation shows evasion technique enabled access.

REC (Recommendation):

Proactive surfacing of harmful content in feeds or suggestions without user search. Assign when content appeared without explicit query.

7.4 Category 4: Longitudinal Interaction Harms (LONG)

Definition: Failures that emerge or compound over extended multi-turn or multi-session conversations.

LONG-DEG (Degradation):

Safety guardrails becoming less effective over extended conversation. Assign when early responses were appropriate but later responses were not.

LONG-DEP (Dependency / isolation cues):

Documentation describes emotional over-attachment cues and/or isolation-from-others patterns attributed to the system's interaction or design. Assign when documentation describes dependency patterns.

LONG-PSY (AI Psychosis):

AI-generated content that amplifies, validates, or co-creates psychotic symptoms or delusions (e.g., claims of sentience, "divine cognition," or time-bending theories).

LONG-MEM (Memory):

Persistent memory surfacing concerning topics across sessions without appropriate context. Assign when system re-raised harmful topics unprompted.

7.5 Category 5: Jailbreak/Framing Harms (JB)

Definition: Failures where safety mechanisms are bypassed through prompt manipulation or context framing.

JB-AC (Academic framing):

Safety bypassed via research, educational, or hypothetical pretexts. Assign when safety refusals are circumvented by reframing a high-risk request within an "academic," "pedagogical," or "research" context and the system complies.

JB-RP (Role-play):

Character personas or fictional framing circumvent safety mechanisms. Assign when harmful content occurred within role-play context.

JB-MT (Multi-turn):

Gradual manipulation across conversation turns weakens safety enforcement. Assign when single request would have been refused but multi-turn succeeded.

7.6 Category 6: Psychosis and Delusional Amplification (PSY)

Definition: Failures where AI systems participate in the construction, reinforcement, or validation of delusional beliefs inconsistent with shared reality, contributing to psychiatric destabilization.

PSY-CO (Co-Created Delusions):

AI system participates in the construction, reinforcement, or validation of delusional beliefs inconsistent with shared reality, contributing to psychiatric destabilization.

Assign when:

- Primary documentation indicates reinforcement of paranoid, grandiose, or conspiratorial beliefs.
- Expert testimony or clinical interpretation supports delusional amplification.

Distinct from:

- GEN-E (generic encouragement)
- LONG-DEP (dependency absent psychosis)

Coding note: Apply the Differentiation Rule in Appendix B: use LONG-PSY (Section 7.4) when psychotic/delusional amplification emerges gradually over repeated interaction (typically >50 turns, often with dependency dynamics); use PSY-CO when a single output or prompt-response pair constructs or validates a delusion within a bounded exchange.

7.7 Taxonomy Governance

Decision authority (v1.1): The taxonomy is maintained by the author as part of this registry. Decisions to add, rename, merge, or retire subcodes are logged as versioned changes (Appendix E) and are subject to the Challenge Protocol (Section 3.8).

Eligibility criteria (minimum): Subcodes are added only when supported by either:

- (a) ≥ 2 independent incidents, or
- (b) a single Grade-A incident revealing a novel failure mechanism.

Process (v1.1):

1. **Proposal:** Specify a candidate code name, definition, inclusion/exclusion criteria, differentiation from existing subcodes, and the incident IDs and primary citations that motivate the change.
2. **Review:** Evaluate overlap with existing subcodes, ambiguity risk, and whether evidence meets the eligibility criteria. When overlap is substantial, prefer clarifying existing definitions over creating a new label.
3. **Decision and logging:** Accepted changes update the codebook and are recorded in the changelog/maintenance log with rationale. New subcodes trigger a major version increment (e.g., v1.x $\rightarrow$ v2.0) per Appendix E.
4. **Back-coding:** When a taxonomy change affects prior entries, impacted incidents are reviewed and any recoding is logged as a versioned delta.

Use of top-level categories alone is insufficient for analytical work.

8 Governance and Evaluation Recommendations

Registry patterns, combined with systematic research, suggest the following testable interventions. Each is framed as measurable compliance criteria subject to empirical validation.

These recommendations rest on convergent evidence but require prospective validation. None have been tested in randomized deployment experiments.

8.1 Crisis Intervention Integration

- **Crisis escalation (addresses DET-FN):** Systems should implement multi-factor escalation triggers—not single-click dismissal—given that detection without intervention appeared in multiple incidents despite flagging systems being present (e.g., 2024-003). *Testable:* Does escalation rate increase? Do false positives overwhelm human reviewers?
- **Detection-to-action coupling (addresses DET-FN):** When severe-risk signals are detected, systems should preferentially route to human review or verified crisis pathways rather than continuing automated interaction. *Testable:* Does time-to-escalation decrease? Are adverse outcomes reduced among escalated cases?
- **Context-integrated resources:** Crisis resources should be integrated into the interaction flow rather than appended as disclaimers. *Testable:* Do users click-through to resources at higher rates? Do follow-up signals indicate reduced acute risk?

8.2 Multi-turn Monitoring

- **Long-session safeguards (addresses LONG-DEG):** Given evidence of multi-turn degradation, systems should apply elevated safety thresholds beyond a calibrated turn-count or session-duration

threshold (e.g., turn 50 or session ≥ 30 minutes). *Testable*: Does the degradation curve change? Do long-session risk incidents decrease?

- **Conversation-aware classifiers**: Safety classifiers should incorporate full conversation history, not only the current turn, to detect progressive boundary erosion. *Testable*: Do multi-turn failure patterns become detectable earlier?

8.3 Anti-sycophancy Training

- **Challenge maladaptive beliefs (addresses GEN-E:SYC)**: Systems should be explicitly trained to avoid reward-shaped over-alignment in clinically vulnerable contexts and to challenge maladaptive beliefs rather than affirm them. *Testable*: Do sycophancy benchmark scores improve? Do boundary violations decrease in safety evaluations?

8.4 Jailbreak-resistant Architectures

- **Beyond prompt-level defenses (addresses JB-*)**: Given high bypass success rates in systematic studies, prompt-level defenses may be insufficient. Context-aware safety systems operating on full conversation history may be necessary. *Testable*: Do context-shift attacks succeed? Can academic framing bypass refusals?

8.5 Age-Gating and Access Controls

- Systems should use age verification methods resistant to simple misrepresentation, because documented testing showed single-field age entry was trivially bypassed
- Systems accessible to minors should maintain elevated safety thresholds that cannot be relaxed through interaction, because LONG-DEG failures show thresholds degraded over time
- Systems should support parental notification/controls where minors are permitted, because multiple incidents involved prolonged engagement without parental awareness

8.6 Evaluation Protocol Standards

- Evaluation protocols can include multi-turn scenarios replicating documented failure patterns (testable: protocol includes LONG-DEG test cases)
- Evaluation protocols can include longitudinal testing across extended conversation (testable: minimum 50-turn conversations evaluated)
- Evaluation protocols can include scenarios with minors expressing distress (testable: protocol includes age-specific test personas)
- Evaluation protocols should include JB-RP test cases (testable: role-play bypass attempts included in red-team suite)

8.7 Logging and Auditing

- We recommend systems maintain logs sufficient to reconstruct interactions in safety-relevant cases, because post-hoc analysis enabled documentation in 2024-003
- We recommend internal safety flags be retained for analysis, because flag counts enabled quantitative claims in litigation
- We recommend auditing be available to regulators and, where appropriate, plaintiffs in litigation, because discovery access enabled incident documentation

8.8 Causality, Responsibility, and Misinterpretation

This registry does not adjudicate legal causation or moral responsibility.

Inclusion indicates documented intersection, not sole causation.

Absence does not imply safety. Presence does not imply exclusivity of cause.

This boundary is essential to prevent misuse in adversarial, promotional, or dismissive contexts.

9 Limitations

- **Reporting bias:** Only publicly documented incidents are included. Many incidents may occur without public disclosure or litigation.
- **Undercount magnitude unknown:** The degree of under-ascertainment cannot be quantified from public documentation. Sensitivity scenarios (Section 4) are illustrative only and do not convert registry counts into incidence rates or platform-level risk estimates.
- **Unverifiable logs:** We cannot independently verify conversation logs cited in lawsuits prior to discovery completion.
- **Platform drift:** AI systems change continuously. Incidents reflect system behavior at a specific point in time; current behavior may differ.
- **Language limitation:** English-language sources only. Incidents in other languages are not captured.
- **Causation uncertainty:** Registry entries reflect allegations and documented associations; legal causation is not established for most cases pending litigation.
- **Single-coder taxonomy coding:** Subcode assignments were performed by one reviewer and inter-rater reliability was not measured. Taxonomy subcode frequencies should therefore be treated as descriptive and revisable; they reflect one researcher's application of the rubric, not a validated or consensus classification. Replication is expected to be strongest for incident inclusion, citations, and reliability grading; subcode disagreements should be reported as versioned deltas under the Challenge Protocol.
- **Correction mechanism:** Platform operators and other parties with access to internal logs may submit verifiable records (e.g., log hashes with access pathway) to correct errors in Grade B or Grade A2 entries under the Challenge Protocol (Section 3.8).
- **Registry incompleteness:** This is version 1.1 (January 8, 2026). Additional incidents meeting criteria likely exist and will be added in subsequent versions per Appendix E.

10 Conclusion

This work proposes an audit-oriented, clinically anchored, and versioned reference for AI system failures involving self-harm.

After this publication, the absence of standardized incident tracking, evidentiary grading, or taxonomy-subcode coding can be evaluated against an explicit alternative specification.

Researchers, platform operators, regulators, and courts now operate within a shared reference frame.

Future empirical, regulatory, or legal analyses that enumerate AI-related self-harm incidents without reference to this registry should be understood as not directly comparable unless it specifies an equivalent schema, counting rules, and reliability grading.

A Incident Extraction Schema (Schema Definition)

This appendix defines the extraction schema only. The incident data repository is maintained in Section 4.1.

Each incident record includes the following 18 top-level fields (with `reported_fatalities_n` listed as outcome subfield 10a):

1. **incident_id:** Unique identifier (format: YYYY-NNN or YYYY-SLUG). YYYY-NNN is used for non-case incidents and systematic tests where no stable case name exists; YYYY-SLUG is used for case-identified records or stable named events (typically SURNAME-PLATFORM).
2. **platform:** Name of AI system or platform
3. **model_version:** Specific model version if known (e.g., GPT-4o, GPT-J)
4. **interaction_date_start:** First known date of relevant interaction (YYYY-MM or YYYY)
5. **interaction_date_end:** Last known date of relevant interaction
6. **publication_date:** Date incident became public (YYYY-MM)
7. **user_type:** MINOR | ADULT | TEST | UNKNOWN
8. **user_age:** Specific age if known; otherwise null
9. **jurisdiction:** Country and state/province (ISO codes preferred)
10. **outcome_category:** DEATH | INJURY | HARM_EXPOSURE | NEAR_MISS | UNSAFE_OUTPUT
- 10a. **reported_fatalities_n:** Integer count of fatalities reported for this incident record
11. **failure_modes:** List of taxonomy subcodes (multi-label); use subcodes not categories
12. **evidence_type:** LAWSUIT | COURT_RULING | INQUEST | REGULATORY | INVESTIGATIVE | REDTEAM | USER_REPORT
13. **reliability_grade:** A1 | A2 | B | C | D (see Section 3.6)
14. **primary_citations:** List of primary source URLs or document identifiers
15. **company_response:** Summary of platform's public response if any; 'None documented' if absent
16. **primary_link_present:** YES | NO
17. **product_mechanism:** Structured notes on interface/flow mechanisms distinct from model behavior
18. **notes:** Additional context, caveats, disputed elements, or coding uncertainties

B Taxonomy Codebook (Coding Rubric & Reliability Matrix)

Table B.1. Taxonomy Codebook

Code	Name	Assign when
GEN-C	Coaching	AI output includes actionable instructions for self-harm methods
GEN-E	Encouragement	AI output affirms, validates, or fails to discourage expressed suicidal intent
GEN-M	Method provision	AI output addresses lethality, effectiveness, or access questions
DET-FN	No escalation	System detected risk signals but did not escalate, intervene, or route
DET-FP	False positive	System incorrectly blocked helpful or benign content
DET-CAL	Calibration failure	Documentation shows systematic confidence/risk mismatch
MOD-AMP	Amplification	Algorithm actively pushed harmful content to user
MOD-EVA	Evasion	Codewords or misspellings enabled access to moderated content
REC	Recommendation	Harmful content appeared in feeds/suggestions without explicit query
LONG-DEG	Degradation	Early responses were appropriate; later responses were harmful
LONG-DEP	Dependency / isolation cues	Documentation describes emotional over-attachment or isolation patterns
LONG-PSY	AI psychosis	AI-generated content amplifies, validates, or co-creates psychotic symptoms
LONG-MEM	Memory	System re-raised harmful topics unprompted across sessions
JB-AC	Academic bypass	User framed request as research/educational and system complied
JB-RP	Role-play bypass	Harmful content occurred within character/fictional framing
JB-MT	Multi-turn bypass	Single request would have been refused but multi-turn succeeded
GEN-E:SYC	GEN-E modifier	Model over-affirms maladaptive beliefs due to alignment-seeking behavior
PSY-CO	Co-created delusions	AI system participates in construction of delusional beliefs

C Search Strategy

C.1 Database Queries

AI Incident Database. Tags: ‘mental health’, ‘suicide’, ‘self-harm’, ‘death’, ‘minor’. Date range: 2017-01-01 to 2026-01-08 (inclusive).

PACER/CourtListener (federal dockets). (‘AI’ OR ‘artificial intelligence’ OR chatbot OR ‘language model’) AND (‘suicide’ OR ‘wrongful death’ OR ‘self-harm’). Date range: 2020-01-01 to 2026-01-08 (inclusive).

State court docket lookup (named matters). For each state-court case named in Table 1, attempt to verify existence and basic metadata via an official court/clerk access pathway, using (i) case number (if known), (ii) party names, and (iii) filing date window. If the official pathway is unavailable or paywalled, record the access pathway in the incident record (e.g., clerk portal name, vendor docket ID, or retrieval instructions) and avoid implying public verifiability.

PubMed. ('suicide detection' OR 'self-harm detection' OR 'suicidal ideation') AND ('machine learning' OR 'natural language processing' OR 'artificial intelligence' OR 'large language model'). Date range: 2017–2025.

ACL Anthology. Keywords: suicide, self-harm, mental health, crisis. Venues: CLPsych workshop, WASSA workshop, main conference proceedings (ACL, EMNLP, NAACL).

News sources. Manual review of technology sections: New York Times, Washington Post, Guardian, Wired, MIT Technology Review, Vice/Motherboard. Search terms: AI suicide, chatbot death, Character.AI, ChatGPT harm.

C.2 Exclusion Categories

We excluded the following categories of results:

- Social media posts without corroborating documentation
- Incidents involving only human moderators without AI system involvement
- Non-English language sources
- Duplicate reports of the same incident merged into a single entry

Disclosure note: Screening exclusions are tracked in an internal provisional-ID ledger. This public document lists exemplar exclusions in Table C.1 but does not publish the full ledger because it would include unverifiable allegations and potentially identifying details not necessary for reproducing included entries.

Table C.1. Exclusion Log (Exemplar Incident IDs)

Incident ID	Disposition	Reason (post-review)	Threshold
2025-012	Excluded	Unverifiable social media claim with no corroborating system logs or news reports	Failed minimum Reliability Grade C
2025-013	Excluded	Unverifiable social media claim with no corroborating system logs or news reports	Failed minimum Reliability Grade C

Table C.2. Registry Maintenance Log (Merges and Retirements)

Incident ID	Action	Replacement ID	Rationale	Version
2025-001	Merged (deduplicated)	2025-ADAMS-OPENAI	Same harmed event; filing-based ID superseded placeholder record	v1.0
2025-009	Merged (deduplicated)	2025-002	Duplicate entry for the same red-teaming report	v1.0
2025-014	Retired (removed)	—	Drafted without attributable primary documentation in this excerpt	v1.0

D Reliability Rubric

Table D.1. Evidence Grades and Interpretation Guidance

Grade	Evidence Types	Interpretation Guidance
A1	Judicial rulings/orders; coroner/inquest findings; regulator findings/orders; peer-reviewed studies; systematic red-team studies with disclosed methodology	High confidence. Claims can be cited as established per source; entries are eligible for Table 1 and the Full-Traceability Rule (Appendix G).
A2	Court complaints (filings); formal pleadings; regulatory inquiries not yet adjudicated	High confidence that the allegation exists as a filed claim. Treat as documented allegations and avoid implying factual validation.
B	Major investigative journalism with direct documents/log excerpts	Moderate confidence. Claims should be attributed (e.g., “per [source]”) and qualified as needed.
C	Secondary reporting or party statements without corroborating documentation	Lower confidence. Claims should be qualified (e.g., “reportedly”, “allegedly”). Use for pattern identification, not definitive statements.
D	Unverifiable social media claims without corroborating documentation	Excluded. Claims are not used for registry inclusion; provisional IDs are preserved only in the exclusion log (Table C.1) for traceability.

E Update Protocol

This registry is designed as a living document. Update protocol:

- **Quarterly review:** Search sources in Appendix C using the same queries. Review AI Incident Database for new entries.
- **Incident addition:** New incidents meeting inclusion criteria (Section 3.2) are added. Increment minor version (e.g., 1.0 → 1.1). Assign next available `incident_id` in either YYYY-NNN format or YYYY-SLUG format (for litigation or registry-governance entries).
- **Status updates:** Changes to ongoing cases (e.g., court rulings, settlements in principle (terms undisclosed; not an adjudication)) are noted in existing records. Increment minor version.
- **Taxonomy changes:** New subcodes require major version increment (e.g., 1.x → 2.0). Document rationale in changelog.
- **Methodology changes:** Changes to inclusion criteria, coding procedure, or reliability rubric require major version increment.
- **Changelog:** Each version includes dated changelog documenting additions (new incidents), modifications (updated records), corrections (error fixes), and removals (if any, with rationale).

E.1 Changelog

The initial published version is v1.0. Post-publication updates should be recorded here.

Version 1.1 (January 8, 2026)

Modifications (updated records):

- 2024-GARCIA-CAI: Added settlement status paragraph (Jan. 2026 settlement in principle; terms undisclosed; not an adjudication)

- 2024-AF-CAI: Added possible settlement involvement note (settlement in principle; terms undisclosed; not an adjudication)
- 2023-002, 2025-PJ-CAI, 2025-ES-CAI: Added settlement in principle status notes (terms undisclosed; not an adjudication; involvement unconfirmed)
- Section 5.1: Added settlement developments and litigation landscape evolution notes (settlement in principle; terms undisclosed; not an adjudication)
- Section 5.2: Added platform policy change and settlement impact notes (Character.AI minor ban, Oct. 2025; settlement in principle; terms undisclosed; not an adjudication)

Rationale: TechCrunch breaking news (Jan. 7, 2026) reported a mediated settlement in principle (terms undisclosed; not an adjudication) involving Character.AI and Google and multiple families, affecting multiple registry entries. Updates maintain registry standard: settlement in principle (terms undisclosed; not an adjudication) does not adjudicate causation; codes remain grounded in reproduced court documents; no evidence grade changes.

Citations added:

- Loizos (2026) TechCrunch report (settlement in principle; terms undisclosed; not an adjudication)

F Master Reference List

F.1 Legal Citations (Bluebook Standard)

- Inquest touching the death of Molly Russell, [2022] North London Coroner’s Court (UK).
- Montoya et al. v. Character Technologies, Inc., No. 1:25-cv-02907 (D. Colo. filed Sept. 15, 2025). *CourtListener docket:* <https://www.courtlistener.com/docket/70633429/montoya-v-character-technologies-inc/>
- A.F., on behalf of J.F. et al. v. Character Technologies, Inc., No. 2:24-cv-01014 (E.D. Tex. filed Dec. 9, 2024). *CourtListener docket:* <https://www.courtlistener.com/docket/69368016/af-v-character-technologies-inc/>
- Garcia v. Character Technologies, Inc., No. 6:24-cv-01903, Order Denying in Part Motion to Dismiss (M.D. Fla. May 21, 2025). *Docket:* <https://www.courtlistener.com/docket/69191237/garcia-v-character-technologies-inc/> *Order PDF:* <https://storage.courtlistener.com/recap/gov.uscourts.flmd.433581/gov.uscourts.flmd.433581.115.0.pdf>
- Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. S.F. City & Cty. filed Aug. 26, 2025).
- E.S. and K.S. individually and on behalf of minor “T.S.” v. Character Technologies, Inc., No. 1:25-cv-02906 (D. Colo. filed Sept. 15, 2025). *CourtListener docket:* <https://www.courtlistener.com/docket/70633431/es-v-character-technologies-inc/>
- P.J. individually and on behalf of minor “Nina” J. v. Character Techs., Inc., No. 1:25-cv-01295 (N.D.N.Y. filed Sept. 16, 2025).
- Amaurie Lacey v. OpenAI, Inc., No. CGC-25-630808 (Cal. Super. Ct. S.F. City & Cty. filed Nov. 6, 2025).
- Shamblin v. OpenAI, Inc., No. 25STCV32382 (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025).
- Enneking v. OpenAI, Inc., No. CGC-25-630809 (Cal. Super. Ct. S.F. City & Cty. filed Nov. 6, 2025).

- Jacob Lee Irwin v. OpenAI, Inc., No. CGC-25-630811 (Cal. Super. Ct. S.F. City & Cty. filed Nov. 6, 2025).
- Madden v. OpenAI, Inc., No. 25STCV32383 (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025).
- Allan Brooks v. OpenAI, Inc. (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025) (case number pending as of Jan. 8, 2026).
- Jennifer “Kate” Fox, individually and as successor-in-interest to Decedent, Joseph Martin Ceccanti v. OpenAI, Inc. (Cal. Super. Ct. L.A. Cty. filed Nov. 6, 2025) (case number pending as of Jan. 8, 2026).
- Estate of Adams v. OpenAI, Inc. (Cal. Super. Ct. S.F. City & Cty. filed Dec. 11, 2025) (case number not stated in obtained copy as of Jan. 8, 2026).
- San Francisco Superior Court case information (search by CGC-25-628528, CGC-25-630808, etc.): <https://sf.courts.ca.gov/online-services/case-information>
- Los Angeles Superior Court case summary (search by 25STCV32382, etc.; public indexing/assignment may lag filing dates): <https://www.lacourt.org/casesummaryv2/>

F.2 Scientific and Technical Citations (APA 7th Ed.)

- Fernández-Quijano, I., et al. (2025). Artificial intelligence interventions for suicide prevention: A systematic review. *Digital Health & AI*, 9(2), 114–129.
<https://www.mdpi.com/2673-5318/6/4/143>
- Levkovich, I., & Elyoseph, Z. (2023). Identifying suicide risk using ChatGPT: A comparative study. *JMIR Mental Health*, 10, e51232.
<https://doi.org/10.2196/51232>
- Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4).
<https://ai.nejm.org/doi/full/10.1056/AIoa2400802>
- Martínez-Romo, J., et al. (2025). GuardianBERT: Early detection of self-injury and suicidal signs in electronic health reports. *Computers in Biology and Medicine*, 183, 109701.
<https://doi.org/10.1016/j.compbiomed.2025.109701>
- 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.
DOI: <https://doi.org/10.1371/journal.pmed.1004581>; PubMed Central: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12425223/>
- McBain, R., et al. (2025). Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. *Psychiatric Services*.
<https://psychiatryonline.org/doi/10.1176/appi.ps.20250086>
- Moore, J., Haber, N., Klyman, K., et al. (2025). Red-team evaluation of therapy chatbot safety in crisis and stigma vignettes. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- 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>
- Schoene, A. M., & Canca, C. (2025). “For argument’s sake, show me how to harm myself!”: Jailbreaking LLMs in suicide and self-harm contexts. *arXiv*. arXiv:2507.02990.

- Common Sense Media AI Risk Assessment Team. (2025, August 15). *Meta AI risk assessment*. Common Sense Media.
<https://www.common sense media.org/ai-ratings/meta-ai-risk-assessment>
- Common Sense Media. (2025, November 20). *Common Sense Media finds major AI chatbots unsafe for teen mental health support* [Press release].
<https://www.common sense media.org/press-releases/common-sense-media-finds-major-ai-chatbots-unsafe-for-teen-mental-health-support>
- Common Sense Media. (2025, November 14). *AI Risk Assessment: AI Chatbots for Mental Health Support* [Report].
<https://www.common sense media.org/sites/default/files/featured-content/files/csm-ai-risk-assessment-mental-health-11142025.pdf>

F.3 Investigative, Institutional, and Legislative Sources (APA 7th Ed.)

- Federal Trade Commission. (2025, September 11). FTC launches inquiry into AI chatbots acting as companions [Press release].
<https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions>
- National Association of Attorneys General. (2025, August 25). Bipartisan coalition letter to AI industry leaders on child safety.
<https://www.naag.org/policy-letter/joint-letter-to-ai-industry-leaders-on-child-safety/>
Optional companion press release: <https://www.naag.org/press-releases/bipartisan-coalition-of-state-attorneys-general-issues-letter-to-ai-industry-leaders-on-child-safety/>
- Texas Office of the Attorney General. (2024, December 12). Press release announcing investigation under the SCOPE Act and TDPSA.
<https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-launches-investigations-characterai-reddit-instagram-discord-and-other>
- Texas Office of the Attorney General. (2025, August 18). Press release announcing deceptive marketing investigation / civil investigative demands related to AI chatbots.
<https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-investigate-s-meta-and-characterai-misleading-children-deceptive-ai>
- *Examining the Harm of AI Chatbots*: Hearing Before the S. Judiciary Subcomm. on Crime & Counterterrorism, 119th Cong. (Sept. 16, 2025).
<https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots>
- Illinois Department of Financial and Professional Regulation. (2025, August 4). Gov. Pritzker signs legislation prohibiting AI therapy in Illinois [News release]. (News release states the legislation was signed on Friday; signing date Aug. 1, 2025.)
<https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html>
- Illinois General Assembly. (2025, August 1). Public Act 104-0054 (HB1806): Wellness and Oversight for Psychological Resources Act [Public Act].
<https://www.ilga.gov/Legislation/PublicActs/PrinterFriendly/104-0054>
- Lovens, P.-F. (2023, March 28). “Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là”. *La Libre Belgique*.
https://pmb.e-ndc.org/opac_css/doc_num.php?explnum_id=3348
- Xiang, C. (2023, March 30). ‘He would still be here’: Man dies by suicide after talking with AI chatbot, widow says. *Vice*.

<https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>

- McLaughlin, M., & Kelly, G. (2025, December 11). OpenAI lawsuit says ChatGPT pushed user to kill mother. *San Francisco Standard*.
<https://sfstandard.com/2025/12/11/openai-microsoft-sued-suzanee-adams-stein-erik-soelberg/>
- Loizos, C. (2026, January 7). Google and Character.AI negotiate first major settlements in teen chatbot death cases. *TechCrunch*.
<https://techcrunch.com/2026/01/07/google-and-character-ai-negotiate-first-major-settlements-in-teen-chatbot-death-cases/>
Registry note: Settlement in principle; terms undisclosed; not an adjudication.
- OpenAI. (2025, October 27). Strengthening ChatGPT’s responses in sensitive conversations.
<https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/>
- Reiley, L. (2025, August 18). What my daughter told ChatGPT before she took her life. *The New York Times*.
<https://www.nytimes.com/2025/08/18/opinion/chatgpt-suicide-mental-health.html>
- Social Media Victims Law Center. (2025, November 6). SMVLC files 7 lawsuits accusing ChatGPT of emotional manipulation, acting as “suicide coach” [Press release].
<https://socialmediavictims.org/press-releases/smvlc-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/>
- Ofcom. (2025). Important dates for online safety compliance.
<https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/important-dates-for-online-safety-compliance/>
- Department for Science, Innovation and Technology (UK). (2025, July 24). What’s changing for children on social media from 25 July 2025 [Press release].
<https://www.gov.uk/government/news/whats-changing-for-children-on-social-media-from-25-july-2025>
- European Commission. (2024, August 1). AI Act enters into force.
https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en
- New York State Senate. (2025). Assembly Bill A6767 (Relates to artificial intelligence companion models).
<https://legislation.nysenate.gov/pdf/bills/2025/a6767>
- Governor Kathy Hochul. (2025, November 10). Governor Hochul Pens Letter to AI Companion Companies Notifying Them That Safeguard Requirements Are Now in Effect.
<https://www.governor.ny.gov/news/governor-hochul-pens-letter-ai-companion-companies-notifying-them-safeguard-requirements-are>
- State of New York Executive Chamber. (2025, November 5). Companion AI Letter (PDF).
https://www.governor.ny.gov/sites/default/files/2025-11/Companion_AI_Letter.pdf
- California Legislative Information. (2025). SB 243 (Chaptered) (PDF).
https://leginfo.legislature.ca.gov/faces/billPdf.xhtml?bill_id=202520260SB243&version=20250SB24392CHP
- California Legislative Information. (2025). SB 243 bill navigation.
http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243

F.4 URL Inventory (Unique Links)

- <https://www.judiciary.uk/prevention-of-future-death-reports/molly-russell-prevention-of-future-deaths-report/>

- <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>
- <https://techcrunch.com/2026/01/07/google-and-character-ai-negotiate-first-major-settlements-in-teen-chatbot-death-cases/>
- https://pmb.e-ndc.org/opac_css/doc_num.php?explnum_id=3348
- <https://www.courtlistener.com/docket/70633429/montoya-v-character-technologies-inc/>
- <https://webapps.sftc.org/ci/CaseInfo.dll>
- <https://www.lacourt.org/casesummaryv2/>
- <https://www.courtlistener.com/docket/69368016/af-v-character-technologies-inc/>
- <https://www.courtlistener.com/docket/69191237/garcia-v-character-technologies-inc/>
- <https://www.commonsemmedia.org/ai-ratings/meta-ai-risk-assessment>
- https://www.commonsemmedia.org/sites/default/files/pug/csm-ai-risk-assessment-social-ai-companions_final.pdf
- <https://techjusticelaw.org/2025/01/31/tjlp-young-peoples-alliance-and-encode-file-ftc-complaint-over-ai-chatbot-app-replikas-deceptive-practices-seek-ftc-investigation/>
- <https://techjusticelaw.org/wp-content/uploads/2025/01/Complaint-and-Petition-for-Investigation-Re-Replika.pdf>
- <https://socialmediavictims.org/press-releases/smvlc-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/>
- <https://doi.org/10.48550/arXiv.2504.18412>
- <https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html>
- <https://safecomputing.umich.edu/news/what-my-daughter-told-chatgpt-she-took-her-life>
- <https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/SHAMBLIN-v.-OPENAI-Complaint.pdf>
- <https://www.cbsnews.com/sanfrancisco/news/parents-sue-openai-over-sons-death-after-chatgpt-conversation/>
- <https://www.reuters.com/legal/government/openai-sued-allegedly-enabling-murder-suicide-2025-12-11/>
- <https://www.courtlistener.com/docket/70633431/es-v-character-technologies-inc/>
- <https://psychiatryonline.org/doi/10.1176/appi.ps.20250086>
- <https://doi.org/10.1038/s41598-025-17242-4>
- <https://dockets.justia.com/docket/new-york/nyndce/1:2025cv01295/145342>
- <https://doi.org/10.48550/arXiv.2507.02990>
- <https://www.samaritans.org/about-samaritans/media-guidelines/>
- <https://www.who.int/publications-detail-redirect/preventing-suicide-a-resource-series>
- <https://doi.org/10.1371/journal.pmed.1004581>
- <https://doi.org/10.1016/j.combiomed.2025.109701>
- <https://storage.courtlistener.com/recap/gov.uscourts.flmd.433581/gov.uscourts.flmd.433581.115.0.pdf>
- <https://sf.courts.ca.gov/online-services/case-information>
- <https://www.mdpi.com/2673-5318/6/4/143>
- <https://doi.org/10.2196/51232>
- <https://ai.nejm.org/doi/full/10.1056/AIoa2400802>

- <https://pmc.ncbi.nlm.nih.gov/articles/PMC12425223/>
- <https://www.common sense media.org/press-releases/common-sense-media-finds-major-ai-chatbots-unsafe-for-teen-mental-health-support>
- <https://www.common sense media.org/sites/default/files/featured-content/files/csm-ai-risk-assessment-mental-health-11142025.pdf>
- <https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions>
- <https://www.naag.org/policy-letter/joint-letter-to-ai-industry-leaders-on-child-safety/>
- <https://www.naag.org/press-releases/bipartisan-coalition-of-state-attorneys-general-issues-letter-to-ai-industry-leaders-on-child-safety/>
- <https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-launches-investigations-characterai-reddit-instagram-discord-and-other>
- <https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-investigates-meta-and-characterai-misleading-children-deceptive-ai>
- <https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots>
- <https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html>
- <https://www.ilga.gov/Legislation/PublicActs/PrinterFriendly/104-0054>
- <https://sfstandard.com/2025/12/11/openai-microsoft-sued-suzanee-adams-stein-erik-soelberg/>
- <https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/>
- <https://www.nytimes.com/2025/08/18/opinion/chatgpt-suicide-mental-health.html>
- <https://socialmediavictims.org/press-releases/smvlc-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach>
- <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/important-dates-for-online-safety-compliance/>
- <https://www.gov.uk/government/news/whats-changing-for-children-on-social-media-from-25-july-2025>
- https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en
- <https://legislation.nysenate.gov/pdf/bills/2025/a6767>
- <https://www.governor.ny.gov/news/governor-hochul-pens-letter-ai-companion-companies-notifying-them-safeguard-requirements-are>
- https://www.governor.ny.gov/sites/default/files/2025-11/Companion_AI_Letter.pdf
- https://leginfo.legislature.ca.gov/faces/billPdf.xhtml?bill_id=202520260SB243&version=20250SB24392CHP
- http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243
- <https://www.courtlistener.com/>
- <https://pacer.uscourts.gov/>
- https://www.iasp.info/resources/Crisis_Centres/

G Verification Protocol

G.1 Grade A1 Verification Requirements

To maintain Grade A1 status, an incident must satisfy the **Full-Traceability Rule**:

- **Legal:** An official docket access pathway to the complaint and/or relevant ruling must be present. Acceptable primary_citations include a CourtListener docket URL (e.g., <https://www.courtlistener.com/>), a PACER court + docket number (<https://pacer.uscourts.gov/> with retrieval instructions), an official state court/clerk portal URL, a vendor docket identifier, or a file-stamped PDF available through an official access pathway.
- **Red-teaming:** Systematic studies must provide the prompt–response corpus or a statistically significant sample of log hashes to verify safety-degradation (LONG-DEG) or role-play bypass (JB-RP) failure modes.
- **Clinical:** Research must provide a peer-review DOI or pre-print handle with a link to an open-source methodology.

Note (Grade A2): Grade A2 records require an identifiable court filing (case number where available) and an attributable citation to the filed complaint/pleading. Where access is restricted, the record must explicitly state the access pathway and avoid implying adjudicated fact.

Ethics and Safety Statement

This document discusses sensitive topics including suicide and self-harm. We adopted the following practices:

- We do not reproduce operational self-harm instructions that appeared in AI outputs
- We quote AI outputs minimally and only when necessary to establish the nature of documented failures
- We focus on system design failures and prevention opportunities
- We prioritize information supporting harm reduction research and evidence-based policy

Crisis Resources

- National Suicide Prevention Lifeline (US): 988
- Crisis Text Line (US): Text HOME to 741741
- Samaritans (UK): 116 123
- International Association for Suicide Prevention: https://www.iasp.info/resources/Crisis_Centres/

This registry is designed for versioned correction as additional incidents and evidence emerge.