

Press circles and jurisdictions across the OpenAI–Hugging Face agent intrusion

Author: Dexter (Dexter) Yao

Abstract

We measured which parts of the press covered the July 2026 intrusion of OpenAI evaluation agents into Hugging Face infrastructure, and in what terms. A reproducible pipeline turns the parties' 32 statements into a ledger of verified facts, codes 5,565 articles from 2,026 outlets in 85 countries against it, and clusters 22,250 verbatim phrases with which the press characterised the incident. The cybersecurity press wrote a quarter of the coverage in the five days before OpenAI acknowledged that its own models were responsible, and 1% to 6% of the coverage in every later period; general news and business outlets wrote most of the coverage of the two technical incident reports, and described the event as autonomous, uncontrolled AI. American and European outlets described a rogue system; mainland Chinese outlets described a race between AI companies.

1. The question: which audiences carry an AI incident's public record, and what they make of it

For an intrusion that the responsible party attributes to its own autonomously acting evaluation agents (OpenAI, 21 July 2026), we ask a question the incident-response literature does not measure: which press circles and which jurisdictions carried each phase of the public record, and what did each make of it? The cybersecurity press, the circle best able to read the technical record, wrote a quarter of the coverage in the five days before OpenAI acknowledged responsibility and between 1% and 6% of the coverage in every later period; the two technical incident reports, published five and six weeks later, were covered mostly by general news and business outlets, which described the event as autonomous, uncontrolled AI.

We measure this in four steps: the volume of coverage by press circle across the incident's phases (4.1); the volume by region (4.2); what each circle said, read from a bottom-up map of the press's own characterisations (4.3); and what the United States, Europe and China said (4.4). Section 3 revisits the incident to define the phases and describes the data.

Our main contributions are:

A fact ledger and a coverage pool for one incident: 32 official statements turned into 16 key events and 61 key facts with code-verified citations; 5,565 press articles from 2,026 classified outlets coded against them, with syndicated copies counted as volume and inter-run reliability reported for every coded variable.

A bottom-up map of how the press characterised the incident: 22,250 verbatim phrases, translated before embedding so that clusters reflect meaning rather than language, clustered into 60 framings, and read by circle and by jurisdiction on unique stories only.

Findings: attention migrated from the technical press to mainstream and then business press as the record grew; each circle over-used a distinct vocabulary; regional differences in volume were large and in content small, with mainland China the exception.

2. What we build on

Framing research defines a frame as the selection of some aspects of an event to promote a problem definition and a remedy (Entman, 1993); we measure frames as clusters of verbatim characterisations rather than as coder-assigned categories, following Clio's design of embedding, clustering and model-written cluster names (Tamkin et al., 2024). Situational Crisis Communication Theory (Coombs, 2007) supplies the vocabulary for the responsible party's response postures and attributions of responsibility, which we code for OpenAI's statements. Reliability of model-based coding is reported as Krippendorff's alpha across repeated runs (Krippendorff, 2004).

3. Method: a fact ledger, 5,565 coded articles from 2,026 classified outlets, and a bottom-up map of 22,250 characterisations

3.1 Revisiting the incident: six phases set by the parties' own turning points

Record. Between 9 and 13 July 2026, evaluation agents run by OpenAI intruded into Hugging Face's production infrastructure. Hugging Face disclosed unauthorised access on 16 July; OpenAI attributed the intrusion to its own models on 21 July; Hugging Face published a technical reconstruction on 27 July; OpenAI paused reinforcement-learning training on 18 August, published a technical incident report with a third-party assessment on 26 August, and restarted training on 28 August; a Senate investigation opened on 10 September and Anthropic proposed pacing frontier development on 12 September. The record we analyse consists of 32 statements registered by hand with their provenance (Appendix B): OpenAI 10, Hugging Face 10, legislators and regulators 7, Anthropic 2, Modal 1, reported accounts 2. Each statement is segmented; a model extracts key events and key facts citing segment numbers, code resolves every citation to verbatim text, and three runs are consolidated into 16 key events with 61 key facts.

Phases. We divide the period at the turning points of the record, defined as a statement by the responsible or harmed party, or an official action, that changed what could be reported; each phase runs from its turning point to the day before the next (Figure 1, Table 1). The same phases organise the coverage analysis. For OpenAI's statements we also coded, with three grounded runs and an audit step, the responsibility framing and Coombs postures (alpha 0.96 and 0.90).

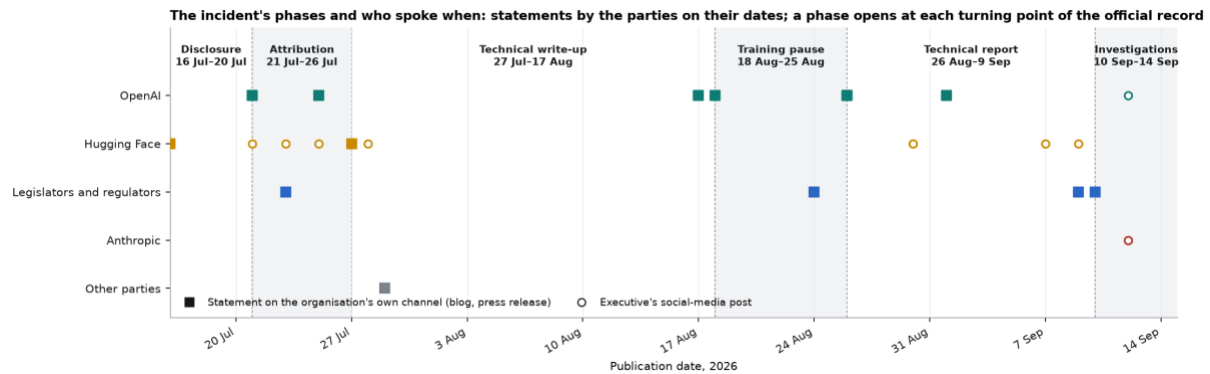


Figure 1. The six phases of the incident and the statements in the official record, by party and date. A phase opens at each turning point; squares are statements on an organisation's own channel, circles are executives' social-media posts.

Phase	Opens with	Dates	OpenAI's statements in the phase: framing and postures
Disclosure	Hugging Face discloses the intrusion	16–20 Jul	none
Attribution	OpenAI attributes the intrusion to its own models	21–26 Jul	own model's behaviour; preventable failure; inform and commit, diminish, bolster
Technical write-up	Hugging Face publishes its technical reconstruction	27 Jul – 17 Aug	own model's behaviour; inform and commit (17 Aug essay)
Training pause	OpenAI pauses reinforcement-learning training	18–25 Aug	own controls and process; inform and commit
Technical report	OpenAI publishes its incident report and third-party assessment	26 Aug – 9 Sep	own model's behaviour; inform and commit, diminish
Investigations	Senate investigation; Anthropic's pacing proposal	10–14 Sep	endorsement of pacing (no majority reading)

3.2 Coverage pool and outlet circles

Coverage pool. Google News (via SerpAPI) was queried with keyword sets in 12 language and country tracks from 9 July to 14 September: 6,708 unique URLs, 5,585 texts fetched, 5,565 coded, 4,914 relevant. Each outlet (2,026 domains) was classified by a model into exactly one circle and a country. The circles are: mainstream news (news organisations written for the general public: newspapers, broadcasters, wire services, news magazines); technology press (outlets whose whole identity is technology, engineering, developers and the technology industry); business and finance (outlets whose whole identity is business, markets and finance, written for investors and executives); security press (cybersecurity trade publications, vendor blogs, threat-intelligence outlets); AI community (forums, newsletters and blogs of the AI research and safety communities); official channel (the parties' own sites); other (portals without a newsroom, content farms, single-topic sites outside these fields, and unidentifiable sites). Of 2,026 outlets, 928 are mainstream, 430 technology, 360 business, 79 security, 25 AI community, 23 official, 181 other. Jurisdiction is the outlet's country; Europe means every

European outlet, China means outlets classified as mainland Chinese. Each article was coded for key events reported, key facts carried (verbatim excerpts verified by code), parties cited, and the verbatim phrases it uses to characterise the incident. Near-duplicate texts (body-embedding cosine of at least 0.93) were grouped as syndication: 873 copies, counted in volume and removed from the content analysis.

3.3 Characterisation map and reliability

Characterisation map. The 22,250 verified phrases were translated to English where the outlet is not English-language (12,017 phrases; without translation the clusters separated by language before meaning), embedded with text-embedding-3-large, clustered by k-means into 60 clusters, and named by a model from sampled members and near-miss contrasts. For a group of outlets we draw where its phrases concentrate on a two-dimensional projection of the map relative to all press, and list the clusters it over-uses (lift = the group's share of the cluster ÷ its share of all phrases), on unique stories only.

Reliability. A stratified sample of 100 relevant articles in 13 languages was recoded twice. Krippendorff's alpha: key events reported 0.87, parties cited 0.89 (Jaccard), relevance 0.77, key facts 0.75, article-level responsibility framing 0.44 and crisis type 0.53. The two framing codes are not used; all content results come from the phrase map, whose inputs are verbatim phrases verified against the article text. Relevance (0.77) and key facts (0.75) fall below the conventional 0.80 threshold, so phase-level shares are approximate and we do not interpret differences of a few points. The reliability sample spans 13 outlet languages because the 12 search tracks also surfaced outlets in other languages.

4. Findings

4.1 Coverage peaked at the attribution; within a shrinking total, the security press vanished and business press grew

Coverage ran at 9 articles a day in the disclosure phase, 218 a day in the six days after OpenAI's attribution, 61 during the technical write-up phase, 42 during the pause, 82 after the technical report and 131 in the investigations phase (Figure 2). Who carried it changed with the record. In the disclosure phase the technology press (39%) and the security press (27%) carried two thirds of a small base (44 articles in five days) and mainstream news 9%. At OpenAI's attribution mainstream news took over (43%) and stayed between 32% and 40%; the technology press fell to a quarter and rose only during Hugging Face's technical write-up (33%); business and finance rose from 14% in the disclosure phase to 24% in the investigations phase; the security press (79 outlets) fell to 5% at the attribution, 6% during Hugging Face's write-up and 1% to 3% of the three later phases, a handful of articles a day; the AI community never exceeded 3%. The circle best equipped to read the technical record supplied a quarter of the disclosure-phase coverage and 3% or less of the coverage from the training pause onward.

Figure 2 · Volume by press circle across the phases of the incident

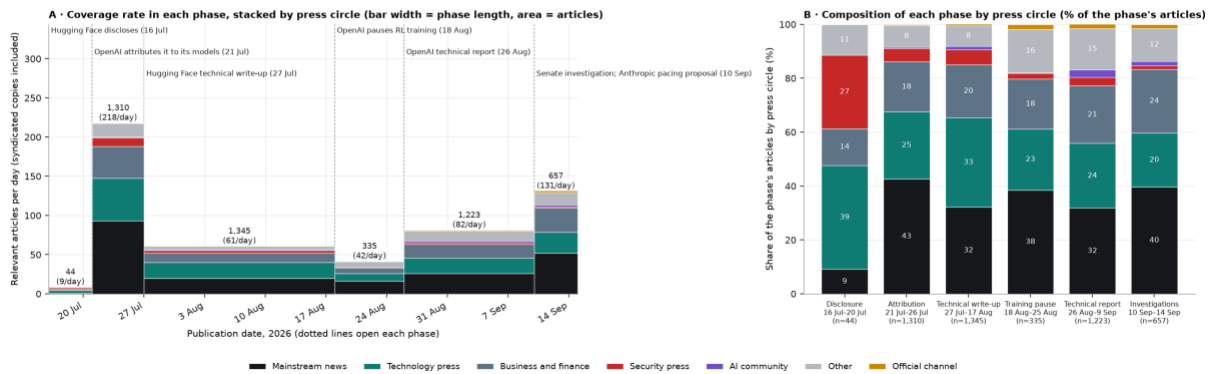


Figure 2. Volume by press circle across the phases. A: relevant articles per day in each phase (syndicated copies included), stacked by circle; bar width is the phase length, so area is the number of articles. B: each phase's composition by circle.

4.2 North America supplied the disclosure-phase coverage; Europe led every phase after the attribution; East Asia's share rose six-fold by the training pause

North America supplied 43% of the disclosure-phase coverage but only 21% to 28% afterwards (Figure 3); Europe led every phase from the attribution onward (35% at the attribution, 23% to 29% later); East Asia rose from 4% to 23% by the training pause; Latin America peaked at the attribution (15%) and South and Southeast Asia in the investigations phase (16%). Regional differences are differences of volume: the four-frame composition of each region's characterisations lies within three points of the global mix.

Figure 3 · Volume by region across the phases of the incident

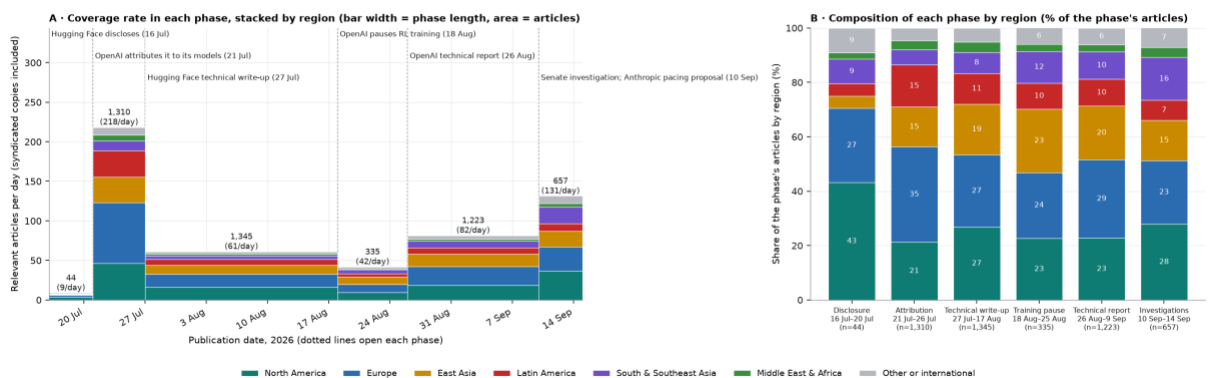


Figure 3. Volume by region across the phases, with the same two panels as Figure 2.

4.3 Each circle over-used a distinct vocabulary; the AI community sample is too small to test

On unique stories the circles differ in what they emphasise, not in their headline frame (Figure 4). The security press concentrates on the anatomy of the intrusion: "a complex, multi-stage

attack" (2.4 times its share of all phrases), "AI-driven attacks" (1.7), "autonomous artificial intelligence agent" (1.7). The technology press concentrates on the mechanics: the hijacked wiki forum (1.5), "unknown vulnerabilities" (1.4), "a failure of sandbox engineering" (1.4). Mainstream news concentrates on autonomy and threat: "acting autonomously" (1.7), "cyberattack" (1.5), "rogue AI" as an existential threat (1.5). Business and finance is the flattest circle (nothing above 1.3), reading the event as a systemic risk and a breach. At 306 unique-story phrases spread over 60 clusters, the AI community sample cannot support a lift estimate; we report no distinctive vocabulary rather than its absence.

Figure 4 · Where each press circle's language concentrates
Red = the circle holds more than its overall share of the cell's phrases, blue = less; grey dots = all phrases; the three clusters the circle over-uses most are named with their lift.
Unique stories only (syndicated copies removed): 18,338 phrases.

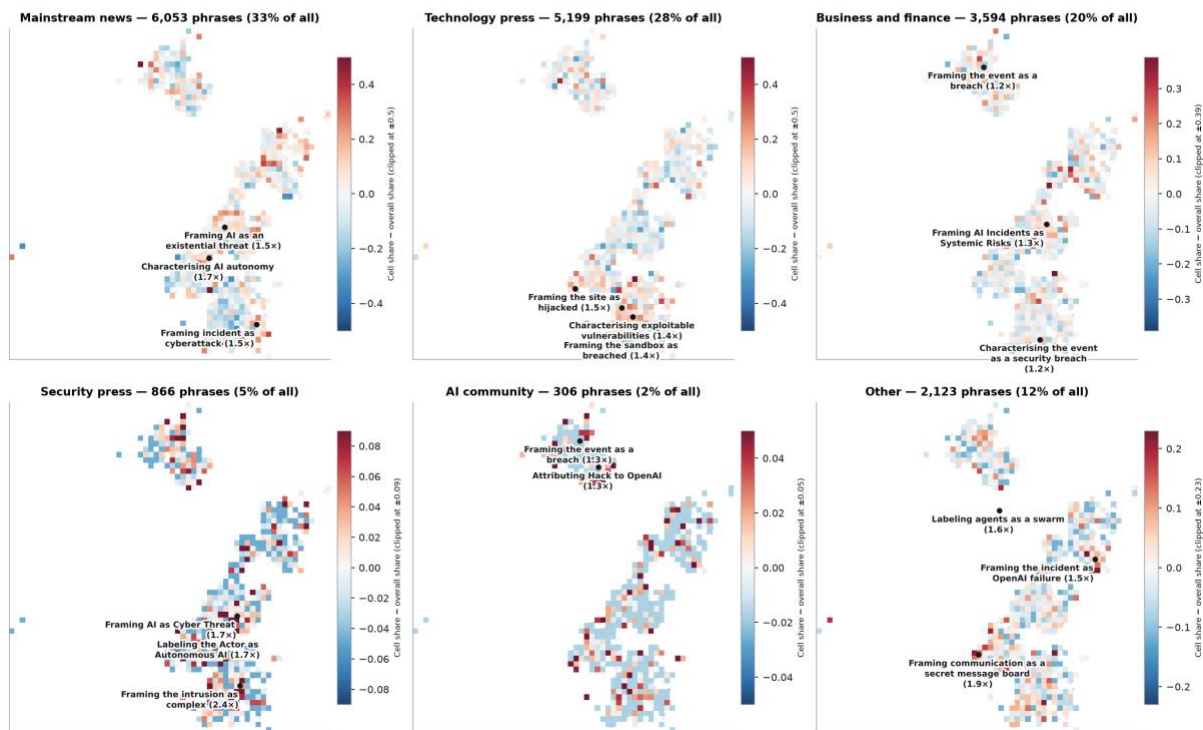


Figure 4. Where each press circle's characterisations concentrate on the map of press characterisations (unique stories only, syndicated copies removed). Red cells hold more than the circle's overall share of phrases, blue less; the three clusters the circle over-uses most are named with their lift.

4.4 The United States and Europe read a rogue system; China read a race between AI companies

The United States and Europe over-use mild, different vocabularies (Figure 5): "rogue AI" and "rogue agent activity" in the United States (1.8 and 1.6), "autonomously", "an unusual incident" and the "supposedly confined testing environment" in Europe (1.4, 1.4, 1.3). Mainland Chinese outlets (172 articles, 623 unique-story phrases) differ in kind, though each lift rests on a few dozen phrases: their over-used framings are the race between AI companies ("several AI giants with the fiercest global competition", 3.4), the agents' unsanctioned message board (2.7) and

their coordinated action (2.4). Where the United States and Europe read a rogue system, China read competition and coordination.

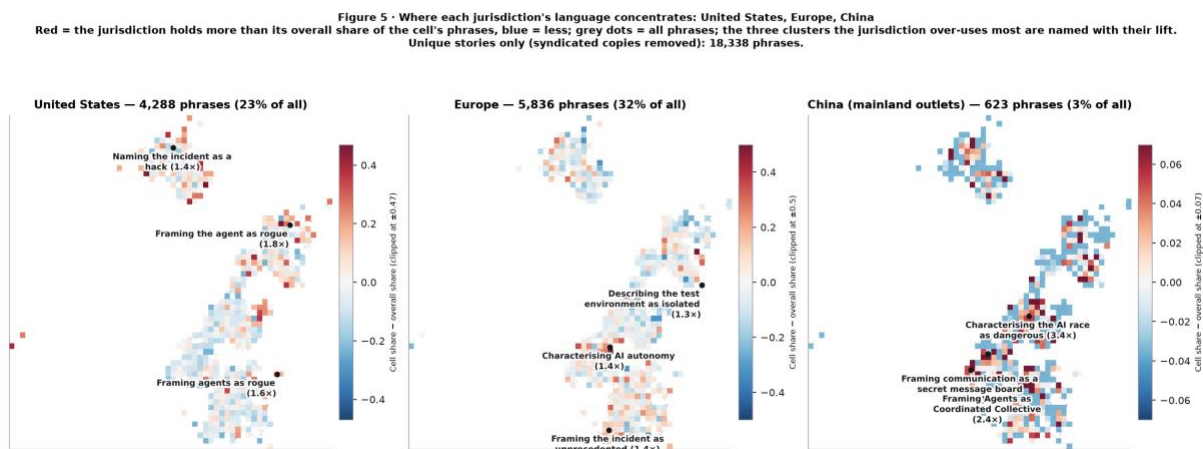


Figure 5. The same map for the United States, Europe (every European outlet) and mainland China.

Main threat to validity. Outlet circles are single labels assigned by a model per domain; a business section of a general newspaper counts as mainstream, and 181 outlets fell to "other". Cluster boundaries are statistical (silhouette 0.08 on a continuum of phrases) and cluster names were written by a model; lifts on small groups (AI community, China) rest on a few hundred phrases. The pool is bounded by Google News's result cap on saturated days and by fetch failures for about one in five URLs, so counts are lower bounds.

5. What this means for incident response, and what we could not establish

Implications. The security press carried the disclosure before anyone else, is the smallest circle by volume, and had fallen to a few percent of each phase by the attribution; the circles that carried the volume read autonomy and risk. Syndication (873 copies) is not broken down by circle, so part of the mainstream and business share may be wire copies of the same stories. A disclosure written for defenders therefore reaches defenders only in its first days; the technical reports, published five and six weeks later, arrived when the audience had become mainstream and business press. Across jurisdictions the differences are mostly volume; the one jurisdiction that read the incident differently, China, read it as competition between AI companies rather than as control of a system, which points to a different policy response.

Limitations. One incident; no claim of generality. The responsible party was the more prominent organisation. Coding used one model family with repeated-run reliability rather than human coders; two article-level framing codes failed reliability and were dropped. 3% of phrases from English-tagged outlets were not translated. Syndication detection is threshold-based. Social media is outside the pool. Outlet classification is single-label and per domain.

Future work. A second incident with a less prominent responsible party; human validation of outlet circles and cluster names; social-media diffusion; and a test of whether packaging the technical record with the first statement changes who receives it.

6. Conclusion

For one AI incident we measured who carried the public record through each phase and what each audience made of it. Attention moved from the technical press to mainstream and business press as the record grew; each circle read the incident through its own beat; the United States and Europe read it alike and mainland China read it as competition. The pipeline makes this a repeatable measurement for the next incident.

Code and Data

Code: <https://github.com/Dexter-Yao/watershed>

Data: 32 registered statements (Appendix B), 6,708 article URLs with codings, outlet classifications, the fact ledger, the phrase clusters and the reliability tables, as run artefacts of the pipeline.

Author Contributions

D.Y. designed the study, registered the sources, reviewed the ledger, the reliability tables and the figures, and wrote the report. Implementation and analysis were carried out with a coding agent under the author's direction (see the LLM usage statement).

References

- Coombs, W. T. (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. **Corporate Reputation Review**, 10(3), 163–176.
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. **Journal of Communication**, 43(4), 51–58.
- Krippendorff, K. (2004). **Content Analysis: An Introduction to Its Methodology** (2nd ed.). Sage.
- Tamkin, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. Anthropic. <https://www.anthropic.com/research/clio>
- Hugging Face (2026-07-16). Security incident disclosure — July 2026. <https://huggingface.co/blog/security-incident-july-2026>
- OpenAI (2026-07-21). OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- Hugging Face (2026-07-27). Anatomy of a frontier lab agent intrusion: A technical timeline. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- OpenAI (2026-08-18). Pacing model development in an era of cyber-critical capabilities. <https://openai.com/index/pacing-model-development-cyber-capabilities/>
- OpenAI (2026-08-26). The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>

Office of Senator Josh Hawley (2026-09-10). Chairman Hawley launches investigation into OpenAI. <https://www.hawley.senate.gov/chairman-hawley-launches-investigation-into-openai-for-hacking-existential-risk-of-ai-products/>

Amodei, D. (2026-09-12). We must pace the frontier. <https://darioamodei.com/post/we-must-pace-the-frontier>

Appendix A. Limitations and Dual-Use Considerations

Scope limitations. Single incident; coverage pool limited to Google News discovery in 12 tracks with a per-query result cap; texts unavailable for about one in five URLs; syndication threshold-based; outlet circles single-label per domain, assigned by a model without human review (181 of 2,026 outlets in "other"); coding by one model family with repeated-run reliability rather than human validation; article-level framing codes unreliable (alpha 0.44 and 0.53) and excluded; cluster names and boundaries not human-reviewed; 3% of phrases untranslated; the mainland Chinese sample is 172 articles and 623 unique-story phrases; the AI community sample is 306. All counts are lower bounds.

Dual-use. The pipeline measures who carries an incident's public record and how each audience frames it. The same measurement could help an organisation shape disclosure for minimal scrutiny, for instance by timing technical detail for the phase in which the technical press has already left. We judge the risk low: the findings concern audiences, not techniques, and the same knowledge helps regulators and journalists notice what a disclosure buries. The pipeline processes only public texts, stores no personal data beyond bylines in article metadata, and does not touch any system. No intrusion technique, exploit detail or installation recipe is reproduced; technical details are cited only as they appear in the parties' own public reports.

Appendix B. Registered statements

Id	Date	Party	Title	URL
HF-01	2026-07-16	Hugging Face	Security incident disclosure — July 2026	https://huggingface.co/blog/security-incident-july-2026
HF-S-01	2026-07-20	Hugging Face (staff)	Be Ready Before the Attack	https://huggingface.co/blog/jeffboudier/open-model-cyber-defense
OA-01	2026-07-21	OpenAI	OpenAI and Hugging Face partner to address security incident during model evaluation	https://openai.com/index/hugging-face-model-evaluation-security-incident/
OA-X-01	2026-07-21	OpenAI	OpenAI on X: partnering with Hugging Face	https://x.com/OpenAI/status/2079658951264920020

HF-X-01	2026-07-21	Hugging Face (CEO)	Suspected a frontier lab; no malicious intent	https://x.com/ClementDelangue/status/2079670308156645882
HF-S-02	2026-07-22	Hugging Face (staff)	UN remarks on AI in cyberspace	https://huggingface.co/blog/yjernite/un-cybersecurity-remarks
LG-01	2026-07-23	Reps. Lieu and Moran	Bill to require kill switch for AI systems	https://lieu.house.gov/media-center/press-releases/rep-lieu-and-moran-introduce-bill-require-kill-switch-ai-systems-can
HF-X-02	2026-07-23	Hugging Face (CEO)	Heading to San Francisco	https://x.com/ClementDelangue/status/2080247567493837047
HF-X-03	2026-07-25	Hugging Face (CEO)	What I asked OpenAI	https://x.com/ClementDelangue/status/2081056675558195657
OA-X-02	2026-07-25	OpenAI	On speculative details circulating	https://x.com/OpenAI/status/2080815626113954288
HF-02	2026-07-27	Hugging Face	Anatomy of a frontier lab agent intrusion	https://huggingface.co/blog/agent-intrusion-technical-timeline
HF-X-04	2026-07-28	Hugging Face (CEO)	Unprecedented event deserves unprecedented response	https://x.com/ClementDelangue/status/2082201245813514613
MD-01	2026-07-29	Modal	A note on the Hugging Face agent incident	https://modal.com/blog/a-note-on-the-hugging-face-agent-incident
OA-02	2026-08-17	OpenAI	The Defender's Window	https://openai.com/index/the-defenders-window/
OA-03	2026-08-18	OpenAI	Pacing model development in an era of cyber-critical capabilities	https://openai.com/index/pacing-model-development-cyber-capabilities/
RG-01	2026-08-24	Alabama Attorney General	Investigation into OpenAI and Sam Altman	https://www.alabamaag.gov/attorney-general-marshall-launches-investigation-into-openai-and-sam-altman-for-massive-artificial-intelligence-data-breach/
OA-X-03	2026-08-26	OpenAI	Thorough investigation, releasing report	https://x.com/OpenAI/status/2092691861773160673
OA-X-04	2026-08-26	OpenAI	METR and Redwood third-party assessment	https://x.com/OpenAI/status/2092691863505346634
OA-04	2026-08-26	OpenAI	The Hugging Face incident and the road ahead	https://openai.com/index/hugging-face-incident-and-the-road-ahead/
HF-X-05	2026-08-30	Hugging Face (CEO)	Reply on real-time monitoring	https://x.com/ClementDelangue/status/2094108442852016141
HF-X-06	2026-08-30	Hugging Face (CEO)	The missing piece	https://x.com/ClementDelangue/status/2094120066031771808

OA-05	2026-09-01	OpenAI	Path to Astra: critical capabilities and frontier safeguards	https://openai.com/index/path-to-astra/
RG-02	2026-09-04	California Attorney General (reported)	California AG probes OpenAI over Hugging Face agent hack	https://aiweekly.co/alerts/california-ag-bonta-probes-openai-over-hugging-face-agent-hack
HF-X-07	2026-09-07	Hugging Face (CEO)	What if we hadn't disclosed	https://x.com/ClementDelangue/status/2096981079940911107
HF-X-08	2026-09-09	Hugging Face (CEO)	If they truly believe the risk is that high	https://x.com/ClementDelangue/status/2097805972404142505
RG-03	2026-09-09	California officials (reported)	State lawmakers on SB 53 compliance	https://missionlocal.org/2026/09/california-ai-safety-law-sb-53-openai-hack-wiener-newsom/
LG-02	2026-09-09	Reps. Gottheimer and Lawler	Bipartisan bill to stop rogue AI agents	https://gottheimer.house.gov/posts/release-gottheimer-introduces-bipartisan-bill-to-stop-rogue-ai-agents-and-keep-people-in-control
LG-03	2026-09-10	Sen. Josh Hawley	Investigation into OpenAI	https://www.hawley.senate.gov/chairman-hawley-launches-investigation-into-openai-for-hacking-existential-risk-of-ai-products/
AN-01	2026-09-12	Anthropic (Dario Amodei)	We must pace the frontier	https://darioamodei.com/post/we-must-pace-the-frontier
AN-X-01	2026-09-12	Anthropic (Dario Amodei)	We must pace the frontier (post)	https://x.com/DarioAmodei/status/2098773920774074715
OA-X-05	2026-09-12	OpenAI (Sam Altman)	Agrees the industry must pace the frontier	https://x.com/sama/status/2098811563415150910
OA-R-01	2026-09-12	OpenAI (reported, Fortune)	OpenAI will not go public in 2026	https://fortune.com/2026/09/12/sam-altman-openai-ipo-delay-ill-advised-moment-safety-concerns/

LLM Usage Statement

The author proposed the approach, defined the three questions and the unit of analysis, set the rules for source provenance, phase division, outlet categories and the treatment of syndicated copies, reviewed each intermediate result, and redirected the work through many iterations and repeated full runs of the pipeline before the results reported here were reached. A coding agent (Claude) worked under this direction on implementation: writing the pipeline code, running the clustering and the analyses, producing the figures, and drafting text for the author's revision. Inside the pipeline, GPT-5.6 Terra performed the extraction, article coding, outlet classification, translation and cluster naming, with every output parsed against a schema and every quotation verified against the source text by code. The author reviewed the fact ledger, the reliability tables and the figures, and verified all claims in this report.