LIMEN: a reviewer-safe public-source observatory for AI edge cases under source, language, and governance uncertainty

Anton Sokolov · Tyche Institute

Preprint draft v0.2 — 2026-06-14 · Methods/data article · Live observatory: https://limen.eatf.eu · Public atlas: https://obscure-ai.eatf.eu

Contents

Abstract . . . 1
1. The problem: edge-case corpora hide their own uncertainty . . . 2
2. Method: gather, verify, tier, bound, link . . . 2
3. The evidence architecture: separate denominator classes . . . 3
4. Demonstration: the reviewed core at scale . . . 4
5. The observatory panels . . . 6
6. Limitations . . . 6
7. Conclusion . . . 7

Abstract

Catalogues of "AI failures" are easy to assemble and hard to trust. They tend to merge a regulator's final order, a charging-stage allegation, a documented software vulnerability, and a viral news anecdote into one undifferentiated count — and then invite readers to treat that count as prevalence. LIMEN takes the opposite stance. It is a public-source observatory whose contribution is an *evidence architecture*: a way of cataloguing AI edge cases that keeps source authority, evidence maturity, duplicate control, legal uncertainty, and language coverage explicit, and that bounds every count by what the underlying record can actually support. We describe the architecture and demonstrate it on a reviewed core of 248 evidence-grade AI edge cases (plus 46 separately-marked media-documented incidents) drawn from regulators, courts, prosecutors and security disclosures across 32 jurisdictions, each adversarially fact-checked against its primary source and bound to it. The reviewed core is presented not as a complete or representative map of AI harm, but as a worked demonstration that an edge-case atlas can grow substantially while remaining reviewer-safe: its denominator classes stay separate, its proof ceilings stay explicit, and no class is asked to support a claim it cannot bear. We make no claim of corpus completeness, incident prevalence, representativeness, legal violation by inference, or a single fused total.

1. The problem: edge-case corpora hide their own uncertainty

Public interest in where artificial intelligence goes wrong has produced a growing number of incident lists and "AI harm" trackers. They are valuable, but they share a structural weakness: the act of putting heterogeneous events in one table quietly equalises them. A binding regulator decision sits in the same column as an unproven complaint; a confirmed software vulnerability sits beside a media anecdote; a single jurisdiction's enforcement action sits beside a global news story. Readers — and especially downstream researchers and journalists — then reach for the row count as if it measured something: how common a harm is, how complete the catalogue is, how one country compares to another. None of those inferences is supported by the way such tables are built.

LIMEN's premise is that the honest contribution of an edge-case observatory is not the size of its corpus but the discipline of its evidence handling. Concretely, an edge-case atlas should make four things visible at all times:

1. **Source authority** — what kind of body produced the record (a court, a regulator, a prosecutor at charging stage, a security coordinator, a news outlet), because that determines what the record can prove.
2. **Evidence maturity** — whether a matter is final, interim, contested or merely alleged.
3. **Duplicate and denominator control** — which counts are distinct, which are views of the same event, and which "totals" are actually view-specific rather than a single corpus total.
4. **Coverage limits** — where language, jurisdiction or access barriers mean absence of evidence is not evidence of absence.

This article describes how LIMEN operationalises those requirements and demonstrates them on a substantial, recently-expanded reviewed core.

2. Method: gather, verify, tier, bound, link

LIMEN's pipeline is deliberately conservative. Each candidate edge case passes through five steps.

Gather. Candidates are drawn from primary public surfaces — regulator decisions and press releases, court dockets and rulings, prosecutor announcements, security advisories and CVE records — and, separately and explicitly labelled, from reputable journalism for incidents that never reached a legal forum.

Adversarially verify. Each candidate is checked against its primary source by an independent, skeptical pass instructed to *refute by default*: to drop a case it cannot confirm exists, and to demote a case whose framing overstates the record (a settlement described as a finding, an investigation described as a penalty, a charging-stage matter described as a verdict, a wrong date or amount, an appeal or reversal omitted). Across the verified legal/regulatory batches, this pass found no fabricated cases; it did remove entries that did not survive scrutiny — for example, an item recorded as a national police enforcement action that, on checking, was a research-and-development project, and an incident that could not be independently confirmed — and it corrected tier and fact errors throughout.

Tier. Each surviving case is assigned to an explicit evidence class (Section 3) rather than to one flat list.

Bound. Each case carries a *claim ceiling* — the strongest statement the record actually supports — and an explicit *caveat* naming what is not proven (appeal status, charging-stage posture, settlement-without-admission, recency risk).

Link. Every case is bound to its primary record — for the large majority (about 93% of evidence-grade cases, and all security and media cases) a direct link, and for the remainder a record locator (authority, date and instrument) sufficient to retrieve the source — so a reader can verify it directly.

The same pipeline drives both the manuscript and the public companion site, where each case is displayed with its tier, claim ceiling, caveat and source link — a test that the discipline survives public presentation, not only internal review.

3. The evidence architecture: separate denominator classes

The core design choice is that LIMEN never reports one corpus total. It reports separate denominator classes, each with its own meaning and its own ceiling. The reviewed core uses four:

Class	What it is	What it supports	What it does *not* support
Regulator / court record	A named regulator decision, court ruling, settlement or official report with a date	The existence of a dated public action	Incident truth beyond the record; prevalence; that the conduct is typical
Contested / interim	A matter opened, charged, filed, or under appeal at a dated stage	That the matter reached that stage	Any finding of liability or guilt; the final outcome
Security disclosure (CVE)	A documented vulnerability in a named AI/agent system, anchored to a CVE	That the vulnerability was disclosed	Exploitation in the wild; real-world harm; prevalence
Notable incident (media)	A widely-reported incident documented by reputable outlets	That a widely-reported event occurred	Any regulator/court finding; legal conclusion; evidence-grade status

The media class is reported but **excluded from the evidence-grade denominator**. It exists because some of the most instructive AI failures — a chatbot taught to be racist within a day, an algorithm that quietly downgraded women's résumés, a home-valuation model whose collapse cost a company hundreds of millions — never produced a legal record, yet are real, well-documented and important to a general audience. Including them while fencing them off from the evidence-grade count is itself a demonstration of the architecture: the catalogue can serve the public without letting media anecdote inflate its evidentiary claims.

4. Demonstration: the reviewed core at scale

The reviewed core currently holds **294** catalogued cases. Of these, **248 are evidence-grade**, distributed across the three evidence-grade classes and kept separate rather than summed (Table 1). A further **46** cases are media-documented incidents, reported but excluded from the evidence-grade denominator.

Table 1. Reviewed-core denominator classes — counts, jurisdiction reach and share of the evidence-grade total. The classes are reported separately and never summed across the evidence-grade boundary; the media class is shown for completeness but excluded from the evidence-grade denominator.

Denominator class	Cases	Jurisdictions	Share of evidence-grade
Regulator / court record	157	27	63%
Contested / interim	80	17	32%
Security disclosure (CVE)	11	1	4%
Evidence-grade total	**248**	**32**	**100%**
Notable incident (media)	46	7	*excluded*

Coverage spans **32 national and supranational jurisdictions**, with a further set of evidence-grade cases coded as global (the security disclosures) or cross-jurisdictional (multinational scholarly-publishing retractions) rather than tied to one country. Evidence-grade cases concentrate in **2016–2026**, the modern era of AI-specific enforcement; the media layer reaches back to **2010** to give historical context (algorithmic-trading flash events, early facial-analysis bias, the first viral chatbot failures). The cases span thirteen themes — legal and procedural contamination, surveillance and biometrics, deepfakes and synthetic media, AI-washing and false capability claims, agentic and control failures, public-sector algorithmic decisions, hiring and education, data-protection enforcement, rights and ethics, finance and insurance, security and agentic vulnerabilities, chatbot safety and harm, and institutional absurdity.

Figure 1 presents the reviewed core as a theme-by-tier composition in which the evidence classes are drawn as separate segments, never as a single bar. The figure is therefore not a ranking of how common each harm is; it is a visibility surface showing how differently the *evidence* is distributed across themes. Security vulnerabilities are entirely CVE-anchored; surveillance and deepfake themes carry a substantial contested/interim share (reflecting charging-stage prosecutions and matters under appeal); legal-and-procedural and data-protection themes are dominated by final regulator/court records.

The reviewed core grew by more than an order of magnitude over a single assembly cycle, from a small curated seed to 248 evidence-grade cases, **without** relaxing any ceiling. That is the intended result: the architecture scales. Growth added breadth (more jurisdictions, more themes, more recent enforcement) but did not convert any class into a prevalence estimate, a completeness claim, or a fused denominator.

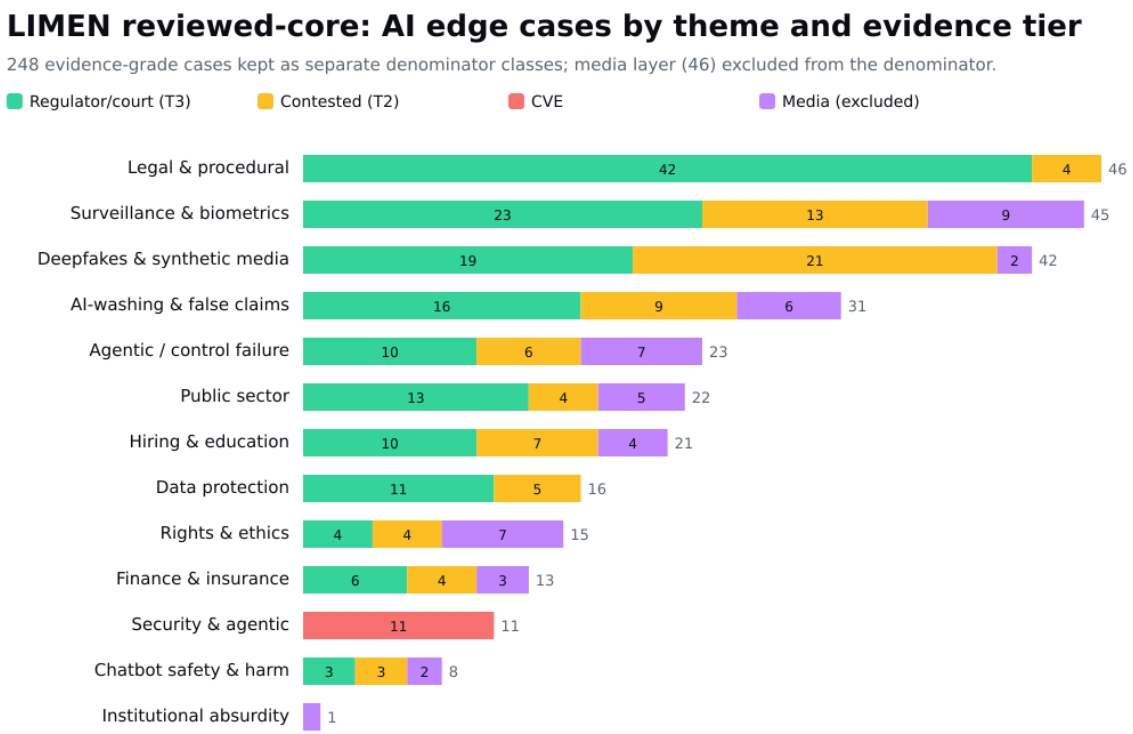

Figure 1: Reviewed core by theme and evidence tier. Each theme's bar is composed of separate evidence-class segments (regulator/court, contested/interim, CVE, media); the segments are never summed into a single magnitude. The chart shows how the *evidence* is distributed across themes, not how common any harm is.

5. The observatory panels

Around the case core, LIMEN maintains a set of analytic panels that apply the same discipline to the wider public-evidence problem. Each panel is denominator-bound and travels with its own claim ceiling; none is interchangeable with another.

- **Source-family map.** Which families of public source are currently usable, which remain thin, and what exact blocker prevents stronger use. It is an observability-boundary surface, not a completeness claim.
- **Taxonomy support heatmap.** Category support over a duplicate-governed denominator, with authority-backed, candidate-only and zero-seed categories kept visibly distinct. Zero-seed categories are a humility channel — they record where the catalogue deliberately withholds a count — not empty cells awaiting a number.
- **Legal-uncertainty matrix.** A hostile-reviewer defence surface that records, per category, how far public evidence can and cannot be pushed toward legal conclusions.
- **Duplicate-control graph.** Join-safety and subtype-overlay infrastructure that keeps distinct records, merged records and overlays separate. It is explicitly *not* recurrence evidence: a denser graph does not mean a harm happened more often.
- **Publication-safe evidence funnel.** The collapse from discovered records to publication-safe lineages, used to discuss evidence maturity and the publication ceiling — not to count incidents.
- **Authority-balance and authority-geography panels.** The composition and geographic concentration of the strongest-anchored records, presented as a limitations-forward companion (a handful of jurisdictions dominate the authoritative tier) rather than as a cross-country comparison.
- **Multilingual visibility map.** Where non-English discovery surfaces local cases that English-centric feeds miss, with explicit translation-review caution on every row. It is an uneven-observability surface, not a country ranking.
- **Security panel.** The agentic-security evidence split into core, appendix-supporting and explicit-gap states, keeping interoperability, remediation depth, source independence and trust-boundary breadth from collapsing into one maturity signal.

The unifying point across panels is the same as for the case core: public evidence about AI edge cases arrives as heterogeneous, partial channels with distinct ceilings, and the observatory's job is to keep those ceilings legible.

These panels, the reviewed case core, and the wider candidate funnel from which the core is drawn (a larger pool of unverified leads that are explicitly *not* counted as cases) are maintained as a live observatory at https://limen.eatf.eu. The public case atlas at https://obscure-ai.eatf.eu is its reader-facing companion: the same reviewed cases, plus the separately-marked media layer, presented for a general audience. The observatory is the source of record for the reviewed core and its proof ceilings; the atlas is the source of the downloadable case dataset.

6. Limitations

LIMEN's limitations are not incidental; they are the boundary that makes the rest defensible.

- **Not complete or representative.** The reviewed core is a catalogue of locatable, source-anchored cases, not a sample of all AI harm. Selection follows evidence availability, which is itself uneven by language, jurisdiction and sector.
- **No prevalence or trend.** No count in this work estimates how often a harm occurs, and the year distribution reflects enforcement and reporting activity, not an underlying rate.
- **No legality by inference.** Contested and charging-stage matters are not findings of liability; security disclosures are not evidence of exploitation; media incidents are not legal conclusions.
- **No fused denominator.** Counts across classes and panels are deliberately not added together.
- **Recency risk.** The most recent (2025–2026) and media-tier items carry higher uncertainty; each links to its source for direct checking.
- **Aggregate geography and recurrence are deferred** until normalised cross-case identifiers exist; the current panels support visibility, not comparison.

7. Conclusion

LIMEN's claim is modest and, we argue, more useful for being modest: a reviewer-safe, evidence-tiered, source-linked observatory of AI edge cases whose denominator classes stay separate and whose claims stay bounded. The reviewed core of 248 evidence-grade cases demonstrates that such an atlas can scale across jurisdictions and themes while holding its proof ceilings, and the public companion site shows the same discipline surviving in front of a general audience. What LIMEN offers is not the biggest list of AI failures, but a defensible architecture for cataloguing them — one in which a reader can always see what each case proves, what it does not, and where to check.

Data and figure artifacts

- Figure 1: `results/dashboard-paper/figure-reviewed-core-tier-by-theme.png` / `.svg`
- Table 1 (evidence panel): `results/dashboard-paper/reviewed-core-evidence-panel-2026-06-14.tsv`
- Tier × theme matrix: `results/dashboard-paper/reviewed-core-tier-category-matrix-2026-06-14.tsv`
- Methods note: `results/dashboard-paper/reviewed-core-corpus-methods-note-2026-06-14.md`
- Live observatory (reviewed core, candidate funnel and analytic panels): https://limen.eatf.eu
- Public case atlas and downloadable dataset: https://obscure-ai.eatf.eu (data at https://obscure-ai.eatf.eu/data/cases.json)

Notes for assembly

- This v0.2 is the reader-facing draft. Internal record identifiers, caption-control codes, boost-lane paths and shard tags from the v0.1 control scaffold (`draft/preprint.md`) are intentionally stripped from reader-facing prose; the scaffold remains the source of per-claim discipline and figure/denominator provenance.

- The observatory-panel denominators (taxonomy core, publication-safe funnel, source families, security split, public-sector and multilingual panels) are governed by the 2026-06-14 dashboard-paper finish packet and must keep their per-panel ceilings when specific counts are quoted.
- For any double-blind venue, the system name and the public-companion URL are de-anonymisation vectors and must be neutralised separately.