EVIWRITE FORMAL REPORT

Training-Data Transparency Reality Check

Whether disclosures are specific, useful, verifiable, actionable, and rightsholder-relevant.

Whether AI training-data disclosures are specific, useful, verifiable, actionable and rightsholder-relevant, with an expanded evidence model for source-to-model reconciliation.

REPORT NUMBER	EW-TDTRC-2026-01
VERSION	1.0
PUBLICATION DATE	2026-05-27
STATUS	Draft for evidence review
DOCUMENT CLASS	Public evidential trend report
REFERENCE	EW-TDTRC-2026-01

REPORT POSITION

Proof landscape, not threat landscape.

EviWrite reports identify where public digital claims become evidentially weak, what stronger records would have required, and how similar claims should be evidenced before pressure arrives.

Document control

Source file	training-data-transparency-reality-check-2026.md
Status	Draft for evidence review
Version	1.0
Period	Evidence horizon: 27 May 2026 (2023-03-14 to 2026-05-27)
Prepared by	EviWrite
Report hash	not-issued
PDF hash	not-issued
Receipt	not-issued

Exclusions

- Legal advice or infringement determinations.
- Forensic audit of any provider, model, dataset or crawler.
- Global incidence statistics for evidence failures.
- Private training-data records not publicly available.

Proof limits

- That any named party committed wrongdoing.
- That any individual case would fail in court or before a regulator.
- That private evidence does not exist.
- That a cited source is complete, final, uncontested, or legally determinative.
- That EviWrite has independently audited the underlying private systems, logs, files, datasets, or forensic records.
- That selected public signal counts represent global incidence.
- That forecast scores represent probability, incidence, legal outcome, or market size.
- That selected international breakdowns represent global prevalence.

Training-Data Transparency Reality Check

Executive summary

Training-data transparency has crossed from policy slogan to evidence problem. The EU now has a public-summary template for general-purpose AI training content; the UK has kept copyright-and-AI transparency under formal policy pressure; the U.S. Copyright Office separates generative-AI training into legally and technically relevant phases; independent transparency indexes still find training data among the most opaque parts of frontier AI; and provenance researchers are showing that authenticity, consent and dataset lineage are infrastructure problems, not public-relations problems.

This report's central finding is narrow and severe: **most training-data transparency still leaves the decisive evidence inside the party being questioned.** A disclosure that cannot be queried, reconciled or challenged is not operational transparency; it is a public-facing inventory of trust. The missing layer is reconciliation. A useful disclosure must connect source, permission basis, opt-out state, collection timing, transformation, training phase, model version and challenge route. Without that bridge, transparency informs the market but preserves the evidential monopoly of the model provider.

This is not a threat-landscape report, legal update, telemetry report, market survey or global incidence study. It is a proof-landscape report. It examines where public AI training-data claims become evidentially weak and what records would make those claims more defensible before dispute, audit, licensing pressure, public challenge or regulatory scrutiny arrives.

Source basis for this report

This report combines official policy sources, technical infrastructure sources, dataset-provenance research, provider disclosures, rights-reservation material, litigation-pressure signals and EviWrite classification. Official sources establish public obligations and policy direction. Technical sources show what evidence could exist at web-crawl, dataset-card, model-card and rights-reservation level. Litigation and stakeholder sources show visible pressure, not legal outcomes.

Source group	Used for	Not used for
Official AI/copyright policy material	Disclosure obligations, policy direction, rightsholder relevance	Determining whether any provider complied or infringed
Dataset-provenance research	Infrastructure weakness, licensing/attribution pressure, documentation limits	Proving any specific work was used by a named model
Web-crawl archive documentation	Capture-level evidence mechanics, WARC/CDX timing, corpus specificity	Proving model-training inclusion
Dataset/model documentation sources	Limits of dataset cards, model cards and metadata claims	Treating documentation as independent custody proof
TDM reservation and opt-out sources	Machine-readable rights signals and future-facing exclusion pressure	Proving that a provider saw, honoured or violated any reservation
Provider disclosures	Public statements and transparency posture	Independent audit of private source registers, training logs or legal basis
Litigation and stakeholder material	Visible dispute pressure and evidential questions	Legal liability, infringement, breach, admissibility or factual findings

Selected-signal boundary: the evidence-signal counts in this report are EviWrite classifications of selected public signals. They are not global incidence, telemetry, proof of wrongdoing, or proof that private evidence does not exist.

Disclosure quality dimensions

The title question of this report is deliberately practical: are disclosures specific, useful, verifiable, actionable and rightsholder-relevant? A disclosure can satisfy public communication expectations while failing all five operational tests.

Disclosure quality	What weak disclosure does	What strong disclosure requires
Specific	Names broad categories such as public web, books, images or code.	Identifies source families, datasets, domains, licences, collection windows and versions.
Useful	Informs policy readers without helping affected actors test relevance.	Lets affected actors assess whether their work, source, domain or licence position may be in scope.
Verifiable	Requires trust in the provider’s private assertion.	Links claims to source registers, versions, logs, manifests, timestamps and preservation records.
Actionable	Provides a static summary or contact page.	Provides challenge, correction, exclusion, decision, appeal and evidence-return routes.
Rightsholder-relevant	Speaks in corpus categories that are too broad for a rightsholder.	Allows work/domain/source-level query and response tied to a model version and evidential boundary.

Evidence basis: European Commission GPAI template and guidance; UK copyright-and-AI report; U.S. Copyright Office generative-AI training report; Stanford FMTI; Data Provenance Initiative; Common Crawl CDXJ

documentation; W3C TDM Reservation Protocol.

Source boundary: This table is an EviWrite evaluation model derived from public sources. It does not determine whether any named provider’s private disclosure process is adequate.

The EviWrite evidencing lens

Every signal is assessed through six questions.

Lens	Question
Source	Where did the record, content, dataset, action or decision come from?
Timing	When did collection, reservation, exclusion, training, release or disclosure occur?
Control	Who controlled the file, system, platform, dataset, crawler, registry, log or model workflow?
Sequence	What happened before, during and after collection, training, disclosure or challenge?
Verification	Can the claim be checked without simply trusting the provider, platform or claimant?
Limits	What does the record not prove?

The EviWrite Evidence Failure Stack

Training-data transparency fails in layers. The public usually sees the top line. The dispute normally lives underneath it.

Layer	Failure question	Common weakness in training-data transparency
Event	Did collection, copying, training, exclusion or challenge occur?	The provider says data categories were used or excluded, but the event is not reconstructable.
Record	Was it recorded at the time?	Evidence is created after regulatory or public pressure begins.
Context	Does the record explain meaning, authority and limits?	A model card, dataset card or public summary exists without source-lineage context.
Custody	Can control and movement be shown?	Provider, platform, crawler or dataset host controls the strongest record.
Trust Boundary	Did the record leave the questioned system?	Evidence remains inside the provider or platform later being challenged.
Verification	Can the claim be checked without belief?	Rightsholders cannot test inclusion, exclusion, permission or model linkage.
Permanence	Will the proof survive time and challenge?	Cards, dashboards, URLs, registries, robots files and policy pages can change.

The Training-Data Transparency Usefulness Test

The minimum public question is no longer “did the provider disclose something?” That is a low bar. The harder question is whether the disclosure lets an affected party do anything.

Test	Evidence question	Weak answer	Evidence-grade answer
Specificity	Can the disclosed source be identified below broad category level?	“Public web data”, “books”, “images”, “code”.	Named corpus, crawl month, dataset version, collection, repository or licensed bundle.
Work relevance	Can a rightsholder test whether their work may be in scope?	No lookup path or only broad categories.	Work-level query route, hashed match, source URL/capture evidence or controlled challenge process.
Permission basis	Is the rights basis stated and evidenced?	“Lawfully accessed”, “public”, “licensed”, or “available online”.	Licence, exception, consent, opt-out state, reservation record and permission evidence linked to source.
Timing	Can collection, reservation, exclusion and training timing be reconstructed?	Disclosure publication date only.	Collection date, crawl date, opt-out capture date, filter date, training start/end and model release date.
Model linkage	Is the data connected to a model version and training phase?	“Used to train our models.”	Model/version manifest mapping source groups to pre-training, fine-tuning, RAG, evaluation or post-training.
Challenge route	Can an affected party dispute or correct the record?	Contact form or policy page.	Ticket, evidence submitted, decision basis, correction/exclusion scope and appeal path.
Evidence return	Does the process produce a durable record?	Email acknowledgement or dashboard state.	Timestamped receipt, source package, decision record and preservation hash.

Evidence basis: European Commission GPAI training-content template; UK Government copyright-and-AI report; U.S. Copyright Office generative-AI training report; Stanford FMTI; Data Provenance Initiative research; Common Crawl CDXJ documentation; W3C TDM Reservation Protocol; Hugging Face model and dataset card documentation.

Source boundary: These sources support the need for a more useful disclosure model. They do not prove that any named provider lacks private evidence.

False confidence patterns

False confidence	Why it is weak	Stronger posture
“We published a training-data summary.”	A summary may satisfy public notice but not enable work-level testing.	Public summary plus confidential source register, query mechanism and challenge receipt.
“We used public web data.”	Public web data is not a source. It is an ocean.	Crawl month, URL/capture index, WARC path, filter manifest and training snapshot.
“The dataset has a card.”	Dataset cards can be incomplete, mutable and disconnected from downstream model use.	Version-pinned dataset manifest, source register, licence fields, archive hash and model linkage.
“The model has a card.”	Model cards can describe behaviour without proving source custody or training phase.	Model-version manifest, training-phase register, evaluation-set boundary and preserved release record.
“Rightsholders can opt out.”	Opt-out without timing, acknowledgement, scope and model-version effect is weak.	Reservation capture, crawler-observation record, provider acknowledgement, exclusion decision and durable receipt.
“We licensed data.”	A licence claim without source-to-work lineage does not tell rightsholders what was covered.	Licence register, covered corpus list, dates, exclusions, downstream model usage and audit trail.
“Open weights are transparent.”	Weight access does not reveal training sources, permissions or opt-out handling.	Open weights plus source-register commitments, data governance record and model lineage evidence.

Five findings

1. Transparency has moved from disclosure to reconciliation

The hard problem is no longer whether providers publish something. It is whether the published statement can be reconciled to sources, permissions, opt-outs, training phases, model versions and rightsholder challenges.

Evidence basis: European Commission GPAI template; UK Government copyright-and-AI report; U.S. Copyright Office generative-AI training report; Data Provenance Initiative research.

Source boundary: These sources support the disclosure-to-reconciliation pressure pattern. They do not prove any private provider record is absent or defective.

2. A disclosure can be specific enough for policy and still useless to a rightsholder

A public category summary can serve a policy function while failing the practical test: can a rightsholder identify, query, challenge, exclude or verify a specific work or source?

Evidence basis: EU public-summary model; UK rightsholder transparency discussion; U.S. Copyright Office analysis of training stages; Stanford FMTI opacity findings.

Source boundary: The sources support a mismatch between public transparency and actionable evidence. They do not establish legal inadequacy in any individual disclosure.

3. The crawl layer is the hidden evidential fault line

Web-scale AI data disputes will increasingly turn on crawl month, capture timestamp, robots or TDM reservation state, crawler identity, filtering and downstream model linkage. Public summaries that omit this layer preserve provider-side evidential control.

Evidence basis: Common Crawl CDXJ documentation; Common Crawl crawl archive notices; W3C TDM Reservation Protocol; WIPO TDMRep material; platform scraping dispute records.

Source boundary: These sources show the evidence mechanics and dispute pressure. They do not prove that any particular model used or misused a specific capture.

4. Model cards and dataset cards are documentation, not custody

Cards help readers understand a dataset or model, but they are mutable explanatory artifacts unless pinned, preserved, complete and linked to source registers, transformations and model-version manifests.

Evidence basis: Hugging Face Dataset Cards documentation; Hugging Face Model Cards documentation; large-scale analysis of Hugging Face dataset cards.

Source boundary: These sources support the distinction between documentation and evidence-grade custody. They do not assess every card or every provider disclosure.

5. Opt-out without receipt is a weak right

Machine-readable rights reservation and Do Not Train tools are important, but rightsholders need proof that the reservation existed before collection, was visible to the crawler, was checked by the provider and was applied to the relevant model version.

Evidence basis: W3C TDM Reservation Protocol; WIPO TDMRep material; Spawning rights-reservation guidance; IPTC opt-out best practices.

Source boundary: These sources support the operational opt-out evidence problem. They do not prove provider receipt, non-receipt, compliance or non-compliance in any named case.

Evidence-signal scorecard

Metric	Count	Meaning
Selected public signals classified	26	Public sources reviewed for evidential relevance.
Signal sectors represented	6	Sectors represented by the selected signal set.
Claim categories represented	6	Controlled claim categories represented in this selected set.
Jurisdiction groups represented	4	UK, EU, North America and international/cross-border signals.
Primary failure types represented	9	Primary evidence weaknesses assigned to the selected signals.
Source records in register	32	Source cards classified by type, source family, use and limitation.

Dataset basis: Derived from the selected evidence signals classified in this report.

Chart boundary: EviWrite classification of selected public signals. Not global incidence, telemetry or legal finding.

Visual chart summary

Charts are generated from the selected public evidence signals and qualitative forecast signals declared in the YAML. They are included to expose classification patterns, not to create fake incidence statistics.

Dataset basis: Derived from the 26 selected evidence signals classified in this report and the six qualitative forecast signals.

Chart boundary: EviWrite classification of selected public signals. Not global incidence, telemetry, legal finding, market share or probability.

Chart	YAML chart ID	Explicit data basis	Boundary
Headline scorecard	chart-scorecard	26 selected signals; 12 failure types; 6 sectors; 4 regions; 6 claim categories; source-family count from source register	Selected signals only
Primary evidence weakness	chart-failure-types	context gap 5; unsupported claim 5; verification gap 5; permission-lineage gap 3; platform-dependent record 3; weak chain of custody 2; fragile metadata 1; late record 1; timing gap 1	Primary failure type only
Sector pressure ranking	chart-sector-pressure	AI and machine learning 10; Copyright and publishing 6; Education and research 3; Platforms and online services 3; Legal and regulatory disputes 2; Media and synthetic content 2	Sector assignment is primary-sector only
Claim-category distribution	chart-claim-category	AI training-data and dataset lineage 16; Platform and third-party record dependency 4; Copyright and authorship claims 3; IP and commercial conflicts 1; Regulatory and operational records 1; Research and education integrity 1	Not global prevalence
Regional distribution	chart-region-distribution	International / cross-border 13; North America 7; European Union 4; United Kingdom 2	Primary public-signal region only
Forward pressure trajectory	chart-forecast-pressure	Six qualitative EviWrite forecast signals with Q1 observed to Q4 forecast scores	Not probability or legal prediction

Visual chart summary

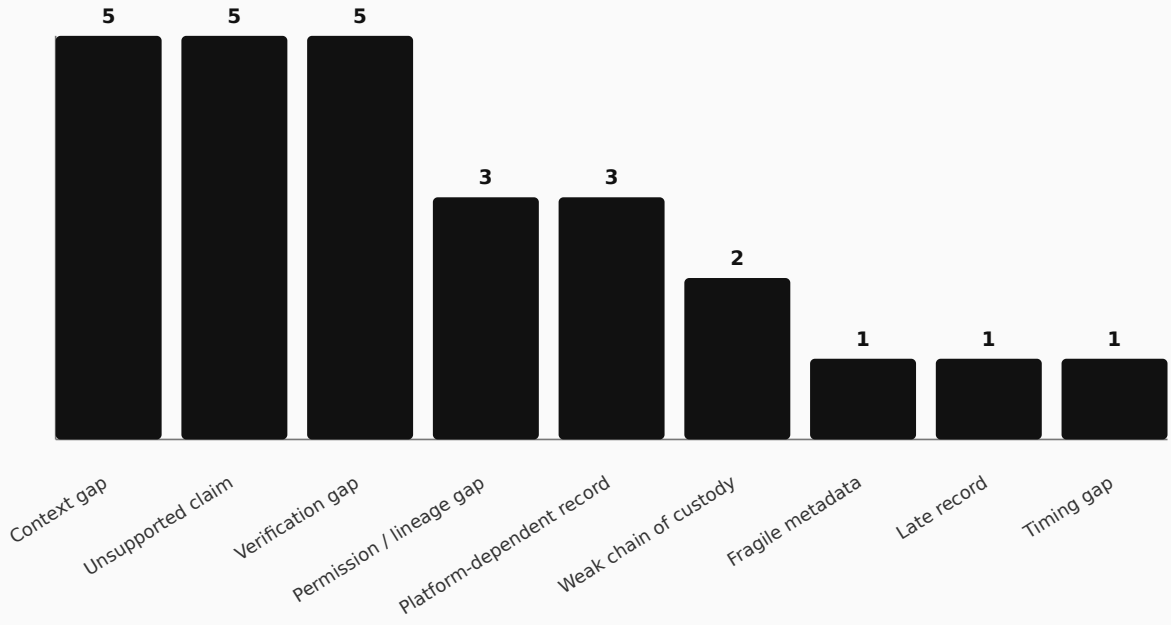
Training-data transparency reality check signals

EviWrite classification of selected public signals. Not global incidence.

26 Public signals classified	**11** Primary and secondary failure types referenced	**6** Signal sectors represented	**4** Region groups represented
6 Claim categories represented	**11** Source families used		

Primary evidence weakness in selected signals

EviWrite classification of selected public signals by primary failure type. Not global incidence.

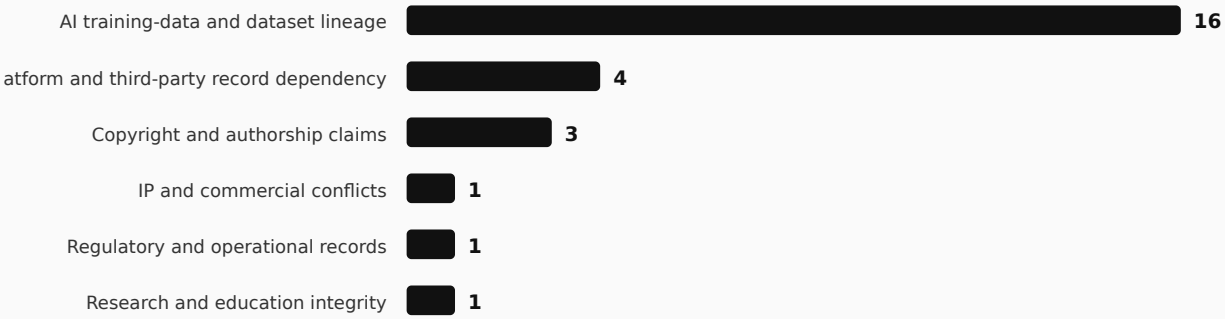


Sector pressure ranking in selected signals

EviWrite classification of selected public signals by primary sector. Not global incidence.

Claim-category distribution in selected signals

EviWrite classification of selected public signals by claim category. Not global incidence.

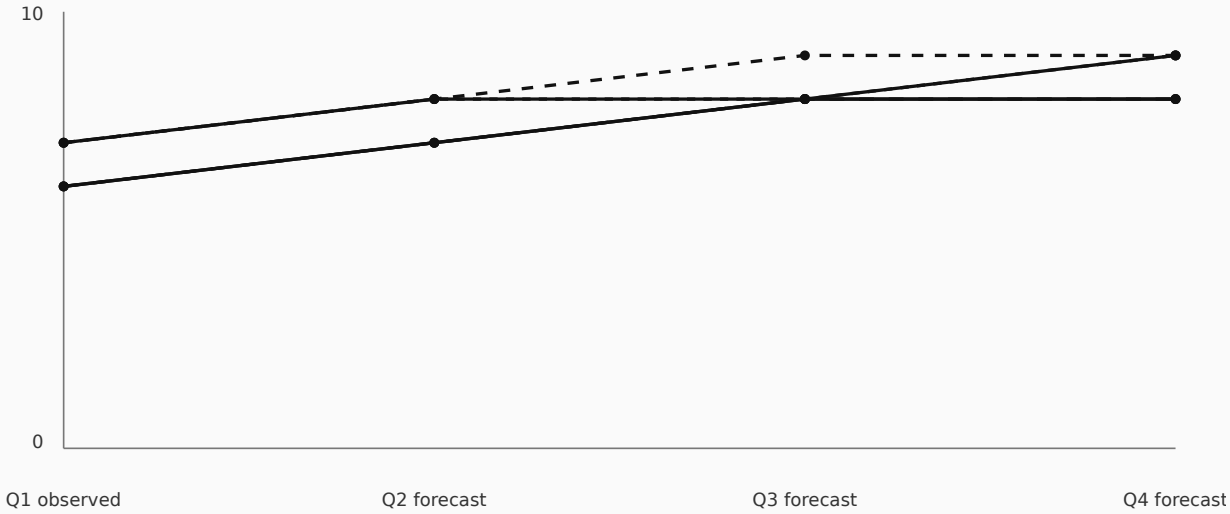


Selected signals by region

EviWrite classification of selected public signals by region. Not global incidence.

Forward evidence-pressure trajectories

Qualitative EviWrite forecast score. Not measured incidence, probability, market sizing, or legal prediction.



1 EU-style public summaries become insufficient witho...

2 UK policy pressure shifts from whether transparency...

3 Web-crawl governance -> a contract and provenance i...

4 Rightsholder query infrastructure -> the next trans...

5 Dataset provenance -> procurement evidence for ente...

6 Synthetic-data claims become a new opacity layer

Evidence failure types by primary weakness

The selected signal set is concentrated around verification gaps, permission-lineage gaps, context gaps, timing gaps and platform-dependent records. That pattern is logical: training-data transparency is usually not defeated by the total absence of any disclosure. It is defeated when the disclosure cannot be tested by the party who needs to rely on it.

Dataset basis: Derived from `evidenceSignals[].primaryFailureType` in the YAML signal register.

Chart boundary: EviWrite classification of selected public signals. Not global incidence, telemetry or legal finding.

Claim-category breakdown

The report is mainly an AI training-data and dataset-lineage report. Copyright/authorship and platform-record dependency appear because they are the pressure channels through which training-data transparency becomes contested. Research and education integrity appears because dataset documentation practices shape the evidence standards that downstream model developers inherit.

Dataset basis: Derived from selected evidence signals classified by controlled claim category.

Chart boundary: This is not a measure of global dispute prevalence.

Sector pressure ranking

The strongest selected pressure sits across AI and machine learning, copyright and publishing, platforms and online services, media and synthetic content, legal/regulatory disputes and education/research. That is the real trajectory: training-data transparency is no longer an AI-policy niche. It is becoming a cross-sector evidential dependency.

Dataset basis: Derived from selected evidence signals classified by sector.

Chart boundary: Sector ranking reflects this report's selected public signals, not global sector incidence.

Evidence-pressure timeline

Date	Signal	Evidence pressure
2024-01	Large-scale Hugging Face dataset-card analysis	Documentation quality becomes measurable rather than assumed.
2024-03	Common Crawl 2024 crawl archive notice	Web-scale corpus specificity exists at archive level but is rarely carried into public AI summaries.
2024-05	W3C TDM Reservation Protocol final report	Rights reservation becomes more machine-readable but still needs receipts and provider acknowledgement.
2024-09	Hamburg Kneschke/LAION decision	Dataset creation, TDM exceptions and opt-out evidence become concrete dispute issues.
2025-05	U.S. Copyright Office Part 3 report	Training is analysed as a sequence of legally relevant acts, not a single vague event.
2025-06	Reddit complaint against Anthropic	Platform data disputes foreground access logs, terms, crawler identity and provider statements.
2025-07	EU GPAI public-summary template	Public training-content summary becomes a structured EU obligation baseline.
2025-12	Stanford FMTI 2025	Training-data transparency remains weak despite rising public pressure.
2026-03	UK Government copyright-and-AI report	UK transparency debate remains active and rightsholder-focused.
2026-05	EviWrite selected-signal review	The practical threshold shifts from disclosure to reconciliation evidence.

Forecast / forward pressure signals

Forecasts are directional EviWrite evidence-pressure assessments based on selected public signals. They are not incidence predictions, legal predictions, probability forecasts or market forecasts.

Forward signal	Expected direction	Evidence question	Minimum Evidence Record
GPAI public summaries become contested	Increase	Is the summary specific enough to support rightsholder action?	Public summary, confidential source register, category-to-source map, challenge route.
Work-level inclusion queries rise	Increase	Can a rightsholder test whether a work was included or excluded?	Work query receipt, source match evidence, capture date, training-phase scope, decision log.
Opt-out moves from signal to receipt	Increase	Can reservation timing and provider acknowledgement be proved?	Reservation capture, crawler-observation log, acknowledgement, scope and model-version exclusion record.
Web-crawl evidence becomes central	Increase	Which capture, crawl, filter and snapshot fed the model?	CDX/WARC references, filter manifests, dataset snapshot hashes, model-version linkage.
Model cards face custody scrutiny	Increase	Does documentation prove lineage or merely describe it?	Version-pinned model card, source-register reference, training-phase manifest and preserved release record.

Source basis: EU GPAI template, UK copyright-and-AI report, U.S. Copyright Office Part 3, Stanford FMTI, Data Provenance Initiative, Common Crawl, W3C TDMRep, Hugging Face documentation, Spawning rights-reservation material.

Forecast boundary: Qualitative evidence-pressure assessment only. Not incidence, probability, legal outcome or market-size prediction.

Additional forward pressure is likely around three under-discussed points. First, rightsholder query infrastructure will become the test of whether transparency is usable. Second, dataset provenance will become procurement evidence for enterprise AI buyers. Third, synthetic-data claims will need lineage records, because synthetic data can inherit unresolved source questions from the model, prompts or seed data that generated it.

Forward pressure signal	Evidence question	Minimum Evidence Record
Rightsholder query infrastructure	Can a rightsholder ask a work-level question and receive a durable bounded answer?	Query intake, submitted-work fingerprint, model scope, search method, response basis, appeal log and receipt.
Dataset provenance in procurement	Can a buyer test source-lineage risk without seeing confidential training data?	Vendor disclosure package, confidential attestation, model inventory, licence/exclusion summary and audit-access clause.
Synthetic-data opacity	Can a provider show what generated the synthetic corpus and what rights/exclusion logic it inherited?	Synthetic generation model/version, seed source register, prompt policy, filter record and synthetic dataset hash.

Forecast boundary: These are qualitative EviWrite evidence-pressure assessments, not incidence predictions, probability forecasts, legal predictions or market forecasts.

Case studies

Case 1: EU GPAI public summary — baseline, not proof

The EU template creates a common public-summary baseline. Its importance is real: it moves transparency from voluntary language toward structured disclosure. Its limit is equally important: a public summary is not a work-level source register, permission register, opt-out ledger or model-version manifest.

Source basis: European Commission GPAI training-content summary template; European Commission GPAI provider guidance.

Evidence weakness: Summary-level disclosure can be too coarse for rightsholder verification.

What stronger evidence would have required:

Weak record	Stronger evidence
Broad data categories	Dataset/source register with corpus, version, date and source-lineage fields
Public summary only	Confidential regulator/auditor register plus public challenge route
No work-level path	Rightsholder query, response, evidence packet and appeal receipt

Evidence boundary: This case study does not assess any provider’s compliance. It examines the evidential boundary of the public-summary model.

Case 2: Common Crawl is not a model-use record

Common Crawl makes web captures and WARC locations queryable. That does not mean a public AI disclosure naming web data or Common Crawl is rightsholder-actionable. The missing bridge is downstream: capture to filtered dataset, filtered dataset to training snapshot, training snapshot to model version.

Source basis: Common Crawl CDXJ documentation; Common Crawl crawl archive notices.

Evidence weakness: Public corpus existence is being mistaken for source-to-model proof.

What stronger evidence would have required:

Weak record	Stronger evidence
“Web data”	Crawl month, URL/capture index, WARC path and timestamp
“Common Crawl”	Filter manifest, deduplication record, excluded-source list and dataset snapshot hash
Provider-held pipeline	Independent preservation of transformation and model-version linkage

Evidence boundary: This does not claim any provider used a specific Common Crawl capture. It identifies the evidence gap.

Case 3: Dataset cards and model cards — context, not custody

Dataset cards and model cards are valuable. They give readers context. But they are not enough as evidence unless pinned, preserved, complete and linked to the underlying records.

Source basis: Hugging Face Dataset Cards documentation; Hugging Face Model Cards documentation; large-scale dataset-card analysis.

Evidence weakness: Documentation can be treated as proof even when it is only an explanatory artifact.

What stronger evidence would have required:

Weak record	Stronger evidence
README-style card	Version-pinned card, revision history, archive hash and source register
General data description	Licence/permission fields, collection dates, transformation records and known exclusions
Model-card training paragraph	Model-version manifest and training-phase source map

Evidence boundary: This case study does not evaluate any specific card’s truthfulness. It assesses the artefact class.

Case 4: Opt-out needs proof of timing and receipt

Machine-readable rights reservation and Do Not Train tools matter because they make rightsholder intent more structured. But the evidential problem remains: did the reservation exist before collection, was it visible, was it checked, was it applied, and to which model version?

Source basis: W3C TDM Reservation Protocol; WIPO TDMRep material; Spawning rights-reservation guidance; IPTC opt-out best practices.

Evidence weakness: A reservation signal is not the same as provider acknowledgement or model-version exclusion proof.

What stronger evidence would have required:

Weak record	Stronger evidence
Policy page or registry entry	Timestamped reservation capture and preservation hash
“Do not train” expression	Provider acknowledgement, scope, model/version applicability and exclusion decision
Dashboard status	Durable receipt and appeal/correction log

Evidence boundary: This does not determine provider compliance or non-compliance. It defines the records needed to test it.

Case 5: Platform data disputes shift proof toward logs and crawler identity

Platform disputes such as Reddit v Anthropic show that training-data transparency is not only about copyright categories. It can become a dispute about terms, access, bots, crawler identity, platform logs, user data controls and provider statements.

Source basis: Reddit docket-stamped complaint; AP reporting on Reddit v Anthropic.

Evidence weakness: The strongest evidence may sit inside platform and provider systems, while affected users and observers see only public allegations and statements.

What stronger evidence would have required:

Weak record	Stronger evidence
Public denial or allegation	Access logs, crawler identity, terms-state record, licence status and preservation package
Output resemblance	Prompt/output evidence, model/version identifier and retrieval/training distinction
Platform-controlled proof	Independent evidential export or court-controlled preservation process

Evidence boundary: Complaint allegations are not findings. This case study is used only for evidence-pressure analysis.

Case 6: Author litigation keeps work-level query pressure unresolved

Author-led disputes show the difference between category-level transparency and work-level usefulness. “Books were in scope” does not answer whether a particular work was copied, retained, transformed, excluded, memorised or used in a model-relevant way.

Source basis: Authors Guild AI class-action overview; U.S. Copyright Office generative-AI training report.

Evidence weakness: Public summaries do not resolve inclusion, copying, fair-use, market-effect or permission questions at work level.

What stronger evidence would have required:

Weak record	Stronger evidence
Training-data category	Work-level query mechanism and source-match evidence
Final model behaviour	Separation of ingestion, training, retention, retrieval and output evidence
Retrospective assertion	Preserved collection and training-phase records created before challenge

Evidence boundary: Stakeholder litigation summaries are visible pressure sources, not proof of infringement.

Additional case signals represented in the YAML

The structured case-study register contains additional compact signals used for charting, source mapping and forecast pressure. They are not expanded into full narrative boxes to keep the body readable.

Additional case signal	Evidential point	Stronger evidence required
Dataset provenance research	Documentation quality varies and often fails source-to-model reconciliation.	Dataset source register, version hash, licence map and transformation record.
Open-source AI definition debate	Openness labels can omit training-data transparency.	Clear separation of weights, code, data information and source lineage.
Kneschke/LAION-style dataset dispute	Public dataset existence does not settle training, permission or model-use questions.	Dataset membership record, exclusion logic, model-linkage evidence and legal-boundary statement.
Publisher licensing announcements	Licence announcements do not reveal non-licensed source treatment.	Licence scope, covered works, excluded works, renewal dates and model-version applicability.
Synthetic-data opacity	Synthetic-data claims can displace rather than solve lineage questions.	Synthetic generation model, seed/input source record, filtering record and licence-inheritance logic.

Evidence boundary: These are selected public evidence-pressure signals, not legal conclusions or incidence measures.

Sector, claim-category and failure-type table

Sector	Claim category	Visible pressure	Primary evidence weakness	Stronger record
AI and machine learning	AI training-data and dataset lineage	GPAI disclosure, model cards, provenance audits	Verification gap	Model-version source register and training-phase manifest
Copyright and publishing	Copyright and authorship claims	Licensing, opt-out and author litigation pressure	Permission-lineage gap	Work-level inclusion query, licence basis and opt-out receipt
Platforms and online services	Platform and third-party record dependency	Scraping, API, terms and crawler disputes	Weak chain of custody	Access logs, crawler identity, terms-state record and preservation package
Education and research	Research and education integrity	Dataset-card practice and reproducibility expectations	Context gap	Version-pinned dataset card plus source register and archive hash
Media and synthetic content	Synthetic media and provenance	Do Not Train and image-data reservation pressure	Platform-dependent record	Rights-reservation capture, acknowledgement and model-version exclusion record
Legal and regulatory disputes	Regulatory and operational records	AI Act, copyright policy and litigation pressure	Timing gap	Event-level collection/training/disclosure chronology

Minimum Evidence Records

A Minimum Evidence Record is the smallest practical record set needed to make a claim more defensible, portable, interpretable and independently checkable. It does not guarantee legal success, regulatory compliance or factual certainty.

Area	Minimum Evidence Record	Why it matters
Public training-data summary	Summary, source-category map, date, model version, update history, limits	Separates public communication from evidence claims.
Confidential source register	Corpus/dataset/source name, version, collection date, licence/permission basis, exclusions, hash or manifest	Gives auditors something testable without exposing every work publicly.
Web crawl data	Crawl month, URL/capture, WARC path, timestamp, crawler identity, robots/TDM state, filter manifest	Converts “web data” into reconstructable evidence.
Dataset documentation	Dataset card, revision history, source register, licence fields, consent/rights fields, archive hash	Prevents documentation from becoming a fragile narrative artifact.
Model versioning	Model identifier, release date, training phase, source snapshot, evaluation data boundary, post-training changes	Stops “model” becoming an ambiguous moving target.
Opt-out / reservation	Reservation method, timestamp, content identifier, crawler visibility, provider acknowledgement, decision, scope	Makes opt-out evidence useful under challenge.
Rightsholder challenge	Submission, evidence provided, provider evidence reviewed, decision, correction/exclusion, appeal, final receipt	Turns transparency into a usable dispute workflow.
Licensed training data	Licence text, covered corpus, work/source list or manifest, term dates, permitted uses, exclusions, audit trail	Distinguishes “licensed data” from vague permission language.
Provider transparency update	Prior disclosure, new disclosure, reason for change, affected model versions, preservation of old version	Prevents silent drift in public claims.

Recommendations by audience

Audience	Action
Creators	Keep original files, publication records, source URLs, rights reservations, opt-out receipts and challenge correspondence. Do not rely on screenshots or platform dashboards alone.
Businesses	Treat AI training-data transparency as a procurement evidence issue. Require source-register summaries, model-version manifests and challenge-resolution records from vendors.
Legal	Ask whether the available record proves source, timing, control, sequence, verification and limits. Avoid treating public summaries as work-level proof.
Providers	Build layered transparency: public summaries, confidential source registers, opt-out receipts, model-version manifests and challenge logs.
AI teams	Record collection, filtering, deduplication, exclusion, training phase, evaluation-set boundary and post-training changes at the time they occur.
Public institutions	Do not procure or deploy systems where the provider cannot explain source-to-model evidence and rightsholder challenge processes.
Education and research	Pin dataset/model cards, preserve dataset revisions, record licences and teach documentation limits as part of AI governance.
Media and publishing	Use machine-readable rights reservations, preserve them independently and require acknowledgement paths from AI developers and platforms.

Methodology and limitations

This report uses selected public sources materially relevant to training-data transparency, rightsholder actionability and evidential readiness. Signals were selected where they exposed a question around source, timing, control, sequence, verification or limits; were relevant to at least one EviWrite audience; and could produce a practical Minimum Evidence Record.

Limitations:

- selected public signals are not global incidence;
- source counts are not telemetry;
- charts are EviWrite classifications unless otherwise stated;
- private provider evidence may exist;
- absence of public evidence is not evidence of absence;
- stakeholder sources are pressure signals, not proof of underlying facts;
- court filings and complaints are allegations unless adjudicated;
- technical standards and documentation sources show possible evidence structures, not provider compliance;
- this report is not legal advice, forensic audit, regulatory finding, assurance engagement or market forecast.

Deep source appendix

The source register is deliberately weighted. The main evidential claims rely on official policy sources, technical infrastructure documentation and independent transparency/provenance research. Litigation records, journalism and stakeholder statements are used to show visible pressure and public dispute signals, not to prove wrongdoing.

Source role	Principal sources	Used for
Primary policy and legal direction	European Commission GPAI material; UK copyright-and-AI report; U.S. Copyright Office Part 3	Public obligations, policy pressure and disclosure limits
Technical evidence mechanics	Common Crawl WARC/CDXJ; W3C TDMRep; Hugging Face card documentation	What evidence could exist and where public artefacts stop
Independent transparency/provenance research	Stanford FMTI; Data Provenance Initiative; dataset-card studies	Pattern support and documentation-quality limits
Pressure and dispute signals	Author litigation, platform-data disputes, Kneschke/LAION, publisher/licensing actions	Visible pressure, not liability findings

Source methodology

Sources are classified by evidential reliability and use, not by whether they support any preferred conclusion. Official sources are used for obligations and policy direction. Academic and institutional datasets are used for pattern support. Technical documentation is used for evidence mechanics. Provider disclosures are used as public statements. Stakeholder and litigation sources are used for visible pressure and evidential questions, not liability findings.

Core repeatable source spine

Source spine	Why it matters here	Main limitation
Official AI/copyright policy material	Establishes disclosure obligations and policy pressure	Does not prove compliance quality.
Court and litigation records	Shows live dispute questions	Allegations are not findings.
Dataset-provenance research	Shows systemic documentation weakness	Does not prove individual model use.
Technical standards and protocols	Shows what evidence could be machine-readable	Adoption and compliance remain separate questions.
Provider transparency material	Shows public disclosure posture	Self-authored, not independent audit.

Expanded source synthesis

The sources reveal a three-layer reality.

First, **policy is moving toward disclosure**. The EU template, UK copyright-and-AI report and U.S. Copyright Office report make training data a governance and legal-policy issue rather than a purely technical issue.

Second, **technical infrastructure can support more precise evidence than public summaries currently expose**. Common Crawl indexes, WARC records, dataset revisions, dataset cards, model cards, metadata standards and TDM reservation protocols show that more granular evidence can exist.

Third, **rightsholder usefulness still depends on reconciliation**. A disclosure that cannot connect source to permission, opt-out timing, model version and challenge outcome leaves the affected party dependent on the provider’s private records.

Selected source register

Source	Type	Used for	Not used for
European Commission - GPAI training-content summary template	Official policy template	EU disclosure baseline and rightsholder-relevance test	Proving provider compliance or work-level use
UK Government - Copyright and AI report	Official government report	UK policy pressure and rightsholder transparency framing	Legal conclusion on any provider
U.S. Copyright Office - Generative AI Training report	Official copyright policy report	Training pipeline phase analysis and copyright-policy context	Determining liability in any case
Stanford CRFM - Foundation Model Transparency Index	Institutional transparency index	Transparency-pattern support and training-data opacity context	Proving any specific work was used
Data Provenance Initiative - dataset licensing and attribution audit	Academic/institutional research	Dataset provenance crisis and licensing/attribution weakness	Provider-specific proof
MIT / Data Provenance research	Academic-policy analysis	Consent/authenticity/provenance infrastructure gap	Legal or forensic findings
Common Crawl - CDXJ index documentation	Technical dataset infrastructure	Capture-level web archive evidence mechanics	Proving model use of a capture
Common Crawl crawl archive notices	Dataset release notice	Crawl timing and scale context	Proving downstream training inclusion
Hugging Face Dataset Cards documentation	Platform documentation	Dataset-card purpose and limits	Proving card completeness or legal basis
Hugging Face Model Cards documentation	Platform documentation	Model-card purpose and limits	Proving source custody
Large-scale analysis of Hugging Face dataset cards	Academic study	Dataset-card completeness and documentation quality	Auditing frontier model training data
W3C TDM Reservation Protocol	Technical protocol	Machine-readable rights reservation	Proving provider receipt or compliance
WIPO TDMRep material	Institutional material	International rights-reservation context	Legal determination
IPTC opt-out best practices	Rights-reservation guidance	Publisher/metadata opt-out practice	Proof of provider acknowledgement
Spawning rights-reservation material	Stakeholder technical guidance	Do Not Train and rightsholder exclusion pressure	Proof a provider honoured an opt-out
Open Source Initiative - Open Source AI Definition	Standards/policy definition	Open model transparency boundary	Training-data disclosure adequacy by itself
MLCommons Croissant	Dataset metadata standard	Structured metadata direction	Proof of use or permission
Kneschke/LAION public decision summary	Court decision summary	Dataset creation and TDM dispute signal	Broad legal conclusions beyond that matter

Source	Type	Used for	Not used for
Reddit complaint against Anthropic	Court filing	Platform-data, crawler and access-log evidence pressure	Liability finding
Authors Guild AI litigation overview	Stakeholder litigation overview	Rightsholder pressure and work-level query problem	Proof of infringement

Source-to-claim map

Claim	Support level	Source basis	Boundary
Training-data transparency is shifting from public disclosure to reconciliation evidence.	Medium-high	EU template; UK report; U.S. Copyright Office; Data Provenance Initiative; Common Crawl; W3C TDMRep	Integrated conclusion is EviWrite interpretation.
Public summaries can inform without enabling rightsholder action.	Strong	EU template; UK report; Stanford FMTI; provider disclosures	Does not prove non-compliance.
Web-crawl evidence requires capture-to-model linkage.	Medium-high	Common Crawl CDXJ; crawl archive notices; TDMRep	Does not prove model use of any capture.
Dataset and model cards are documentation, not custody.	Medium-high	Hugging Face documentation; dataset-card study	Does not assess every card.
Opt-out needs durable receipt evidence.	Medium-high	W3C TDMRep; WIPO TDMRep; Spawning; IPTC	Does not prove provider receipt or non-receipt.
Litigation pressure shows work-level query demand.	Medium	U.S. Copyright Office; Authors Guild; Reddit complaint; Kneschke/LAION	Allegations and stakeholder views are not findings.

Assurance and review status

This is a public evidential trend report. It is not an audit, legal opinion, forensic report, regulatory finding, assurance engagement, cyber telemetry report or global incidence report. The report includes a source register, source-to-claim map, selected-signal register, forecast-signal register, chart limitations, proof limits and body-level source support.

Version and audit record

Field	Value
Report title	Training-Data Transparency Reality Check
Version	2.0
Prepared by	EviWrite
Status	Draft
Publication date	2026-05-27
Evidence horizon	2026-05-27
Report hash	Not issued
PDF hash	Not issued
Source register hash	Not issued
Machine-readable register hash	Not applicable
Receipt	Not issued for this version

Glossary

Term	Definition
Training-data transparency	Disclosure or evidence concerning data used to train, fine-tune, evaluate, align or otherwise develop AI models.
Source register	A structured record identifying sources, datasets, permissions, dates, transformations and model-linkage evidence.
Rightsholder-actionable	Useful enough for a rightsholder to identify, query, challenge, exclude or verify a relevant work or source.
Reconciliation layer	The evidence bridge connecting public disclosure to source, permission, timing, transformation, training phase and model version.
WARC/CDX	Web archive and index formats used to preserve and locate web captures.
TDM reservation	Machine-readable expression that text-and-data-mining rights are reserved or subject to policy.