EVIWRITE FORMAL REPORT

AI Training Data Disputes Monitor

Lawsuits, settlements, licensing disputes, opt-out conflicts, dataset claims, publisher actions, and rightsholder pressure

An EviWrite proof-landscape report on lawsuits, settlements, licensing disputes, opt-out conflicts, dataset claims, publisher actions, and rightsholder pressure around AI training data.

REPORT NUMBER	AITDDM-2026-05
VERSION	1.0
PUBLICATION DATE	not-issued
STATUS	draft-source-reviewed
DOCUMENT CLASS	Public evidential trend report
REFERENCE	AITDDM-2026-05

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	ai-training-data-disputes-monitor.md
Status	draft-source-reviewed
Version	1.0
Period	Public signals to 27 May 2026 (2023-07-13 to 2026-05-27)
Prepared by	EviWrite
Report hash	not-issued
PDF hash	not-issued
Receipt	not-issued

Exclusions

- Private discovery records, sealed filings, confidential licence terms, private datasets, forensic model inspections, non-public logs, and legal determinations not made by competent sources.
- Global incidence measurement, market sizing, legal advice, liability findings, or forensic conclusions.

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.

AI Training Data Disputes Monitor

Executive summary

AI training-data disputes have moved beyond the question most public debate still repeats: whether copyrighted or platform-hosted material was used. The harder evidential question is narrower and more durable: can a specific work be connected to a specific source path, permission basis, crawler event, dataset state, model version, output, exclusion, settlement claim, or licence scope?

This is not a legal update, market map, telemetry report, or global incidence study. It is a proof-landscape report. It examines selected public signals where AI training-data claims become difficult to support because records are missing, late, captive, ambiguous, fragile, or too broad for the claim placed on them.

The selected signals reviewed here point to five pressure points: work-level lineage, licence-scope evidence, opt-out proof, crawler logs, and settlement/remediation records. The mistake is to treat these as legal-administration details. They are now core evidence infrastructure.

Source basis for this report

This report combines repeatable copyright-AI source families with selected public records available by 27 May 2026. EviWrite classification is applied to identify proof weakness. Selected signal counts are not global incidence.

This first monitor issue establishes a baseline selected-signal register through 27 May 2026; future issues may use shorter periodic windows.

Source group	Used for	Not used for
Official copyright and AI policy material	Training-data transparency, copyright-policy pressure, and evidence design	Determining compliance or infringement by any named party
Court records and dockets	Public procedural posture, claim structure, and evidentiary pressure	Treating allegations as findings
Settlement administration records	Claim-process, remediation, and covered-work evidence pressure	Proving liability or full private remediation history
Licensing announcements	Visible content-licensing and platform-data market pressure	Proving full contract terms or licence sufficiency
Opt-out and crawler-control sources	Reservation, access, crawler, and platform-control evidence	Proving enforceability or compliance in any individual case
Journalism	Public-signal discovery and context where primary records are not included	Legal conclusion, forensic truth, or private evidence absence

Legal posture is fragmented. Thomson Reuters v Ross, Kadrey v Meta, Anthropic-related proceedings, OpenAI MDL activity, and publisher actions against Meta do not point to one clean legal outcome. They point to a record-design

problem: different theories require different evidence, including source acquisition, market substitution, training purpose, output similarity, licence scope, copyright-management information, remediation, and model-version linkage.

The EviWrite evidencing lens

Every signal in this report is assessed through six questions:

Lens	Question
Source	Where did the record, content, dataset, action, or decision come from?
Timing	When did the relevant event, creation, access, disclosure, decision, or knowledge occur?
Control	Who controlled the file, system, account, platform, dataset, log, or workflow?
Sequence	What happened before, during, and after the claim being made?
Verification	Can the claim be checked without simply trusting the claimant or the platform?
Limits	What does the record not prove?

The EviWrite Evidence Failure Stack

Layer	Failure question	Common weakness
Event	Did the thing happen?	Access, copying, training, exclusion, deletion, or output is asserted but not reconstructed.
Record	Was it recorded at the time?	Evidence is created after dispute, lawsuit, licensing pressure, or public scrutiny begins.
Context	Does the record explain meaning, authority, and limits?	A transparency summary, licence announcement, output, or opt-out exists without enough context.
Custody	Can control and movement be shown?	Dataset, crawler, platform, vendor, or provider controls the strongest record.
Trust Boundary	Did the record leave the system or party that may later be questioned?	Evidence remains inside the AI system, platform, or publisher workflow being questioned.
Verification	Can the claim be checked without belief?	The public claim depends on private logs, confidential terms, or unsupported assertion.
Permanence	Will the proof survive time and challenge?	Crawler logs, source URLs, opt-outs, manifests, and platform records may change or disappear.

False confidence patterns

False confidence	Why it is weak	Stronger posture
We published a transparency summary	A summary may describe content categories without proving whether a named work was used.	Summary plus source register, dataset manifest, licence basis, and named-work verification route.
We have a licence announcement	Public announcements rarely disclose schedules, exclusions, model coverage, audit rights, or downstream use.	Licence schedule, authority record, permitted-use map, model/version coverage, audit and revocation trail.
We put an AI opt-out online	A current opt-out does not prove historical state, crawler receipt, or compliance.	Timestamped snapshots, protocol version, server logs, crawler identity, and preservation record.
The case settled	Settlement can resolve claims without reconstructing all historical training facts.	Claim files, work identifiers, source status, deletion/exclusion logs, and settlement-scope boundaries.
The platform has logs	The strongest access record may be controlled by the party being questioned.	Independent exports, hash-linked log preservation, retention terms, and access-evidence clauses.

Five findings

1. The training-data dispute has become a work-level lineage dispute

The durable question is no longer whether AI systems train on broad content categories. It is whether a specific work can be connected to source, permission, access, copying, transformation, retention, exclusion, output, and model version.

Evidence basis: U.S. Copyright Office and UK Government reports support copyright/training transparency pressure; EU GPAI materials support public-summary and provider-record pressure; OpenAI MDL material supports continuing litigation pressure around source-lineage allegations.
Source boundary: These sources support governance-record and work-level transparency pressure. They do not establish infringement, fair use, private dataset contents, model-level inclusion, or adequacy of any party’s internal records.

2. Licensing is becoming evidence infrastructure

AI content licensing now needs evidence of source schedules, exclusions, duration, model coverage, downstream use, renewal, audit rights, and revocation/remediation handling. A licence headline is not enough.

Evidence basis: News Corp/OpenAI and AP/OpenAI announcements support visible licensing activity; Reuters Reddit/Google reporting supports platform-data licensing pressure.
Source boundary: These sources support visible licensing-market pressure and selected examples. They do not disclose full contract terms, exclusivity, audit rights, model coverage, or licence sufficiency.

3. Opt-out is becoming a timestamped proof problem

Machine-readable reservation mechanisms matter, but they do not by themselves prove when a reservation existed, whether a crawler received it, whether access occurred before it, or whether downstream systems respected it. Opt-out systems can create a second-order evidence problem: the rightsholder must prove not only the reservation, but the technical exposure of the reservation to the relevant crawler at the relevant time.

Evidence basis: IPTC and EDRLab sources support machine-readable opt-out and rights-reservation practice; Cloudflare material supports crawler-control and access-response pressure.
Source boundary: These sources support the evidence design problem around reservation and crawler access. They do not prove enforceability, crawler receipt, compliance, or non-compliance in any individual case.

4. Settlements do not close the evidence gap; they expose it

A settlement can resolve claims procedurally or commercially while leaving the deeper evidential problem intact: which works, which copies, which systems, which models, which deletion or exclusion actions, and which future uses are covered.

Evidence basis: Reuters Anthropic settlement report; Anthropic settlement administration site.
Source boundary: These sources support public settlement, claims-process, and work-identification analysis. They do not establish liability, infringement, exact private dataset contents, complete remediation history, or final class distribution outcomes.

5. Crawler monetisation is turning access logs into commercial evidence

As crawler access becomes blockable, licensable, or chargeable, logs become economic and legal evidence: who accessed, what was requested, what terms applied, what response was served, and whether content was later used for training or retrieval.

Evidence basis: Cloudflare and Stack Overflow sources support crawler-control and paid-access mechanics; Reddit/Anthropic and Reddit/Google public signals support commercial evidence pressure around platform-controlled content access.
Source boundary: These sources support public crawler-control and monetisation pressure. They do not prove any individual crawler’s compliance, model use, payment obligation, or contractual breach.

Evidence-signal scorecard

Metric	Count	Meaning
Selected public signals classified	13	Official, court, settlement, platform, licensing, opt-out, and public-reporting signals reviewed.
Primary failure types represented	9	Distinct primary weaknesses across the selected signal set.
Signal sectors represented	4	Distinct sectors represented by evidenceSignals[].sector.
Region groups represented	4	Distinct region groups represented by evidenceSignals[].region.
Claim categories represented	4	Distinct claim categories represented by evidenceSignals[].claimCategory.
Source records used	22	Human-readable source register entries supporting body claims and signal classification.

Charts and counts show EviWrite classification of selected public signals. They are not measures of global incidence.

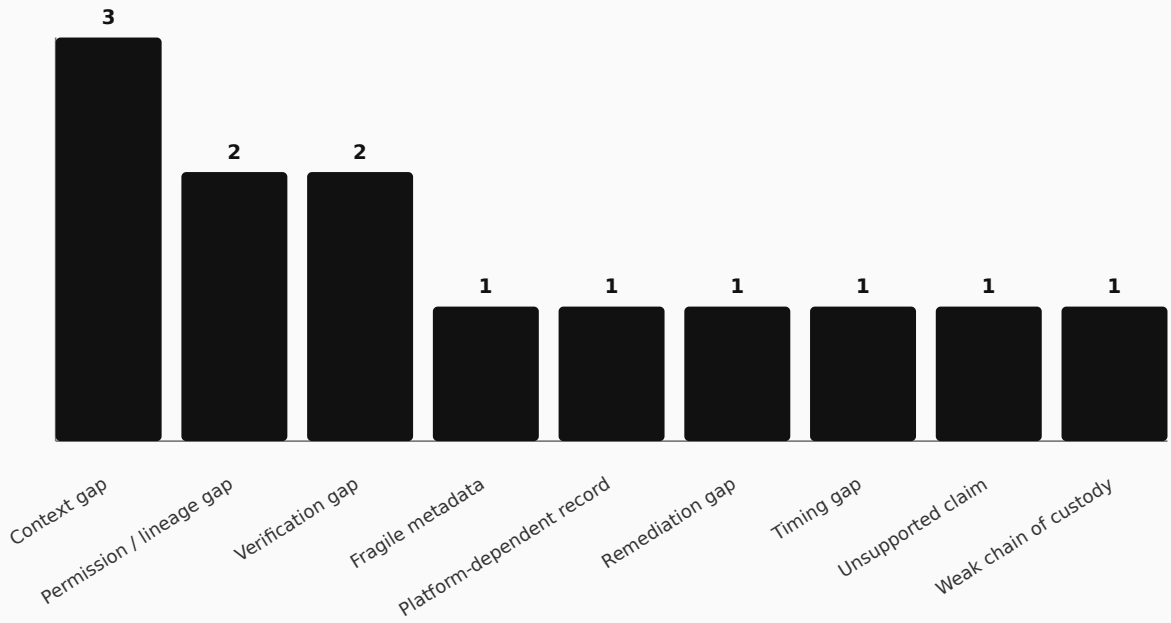
Visual chart summary

Dataset basis: Derived from the selected evidence signals classified in this report.
Chart boundary: This is EviWrite classification of selected public signals. It is not global incidence, telemetry, or legal finding.

Visual chart summary

Evidence failure types by primary weakness

EviWrite classification of selected public signals. Not global incidence.

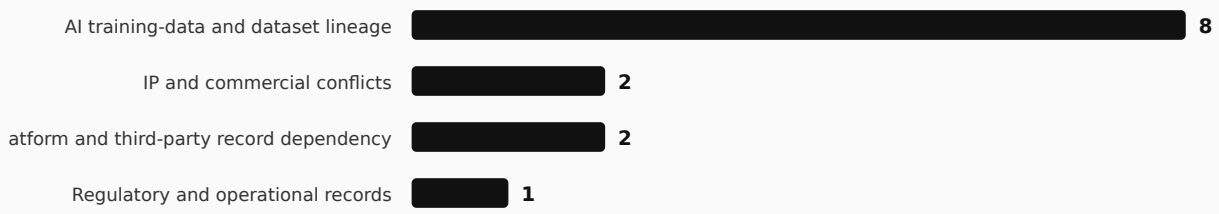


Sectors with visible AI training-data dispute pressure

EviWrite classification of selected public signals. Not global incidence.

Selected signals by claim category

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

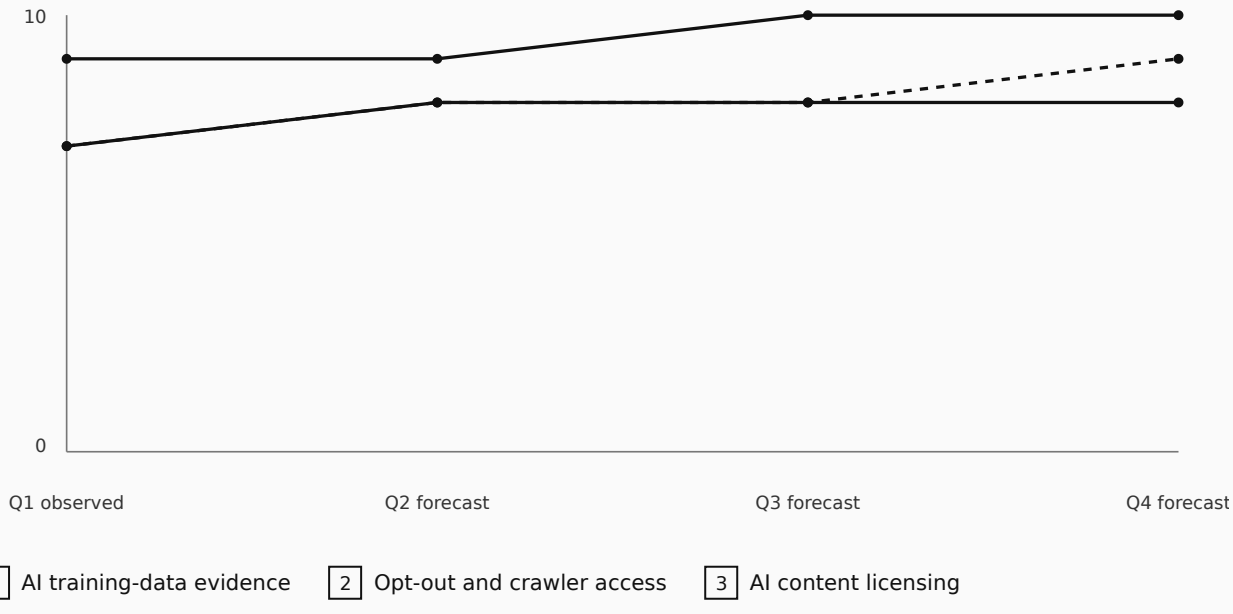


Selected signals by region

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.



Evidence failure types by primary weakness

Primary weakness	Selected signals
context-gap	3
permission-lineage-gap	2
verification-gap	2
remediation-gap	1
unsupported-claim	1
weak-chain-of-custody	1
platform-dependent-record	1
fragile-metadata	1
timing-gap	1

The strongest pattern is not mere absence. It is mismatch: a record exists, but it is too broad, too late, too private, or too detached from the specific work, crawler event, licence, model version, or output claim.

Claim-category breakdown

Claim category	Selected signals
ai-training-data	8
ip-commercial-conflict	2
platform-record-dependency	2
regulatory-operational-records	1

Region distribution

Region	Selected signals	Boundary
North America	9	Selected public signals only
International / cross-border	2	Selected public signals only
European Union	1	Selected public signals only
United Kingdom	1	Selected public signals only

Sector pressure ranking

Sector	Selected signals
Copyright and publishing	5
Legal and regulatory disputes	3
AI and machine learning	3
Platforms and online services	2

Evidence-pressure timeline to 27 May 2026

Date	Signal	Primary evidential pressure
2023-07-13	AP/OpenAI content collaboration	Licence scope and archive-use evidence
2024-02-22	Reddit/Google licensing report	Platform-data licensing and source-control evidence
2025-02-11	Thomson Reuters v Ross order	Structured content, market context, and fair-use evidence
2025-04-11	OpenAI copyright MDL docket signal	Consolidated source-lineage and discovery pressure
2025-05-09	U.S. Copyright Office training report	Source, permission, market, and policy evidence
2025-05-28	IPTC generative-AI opt-out guidance	Timestamped rights-reservation and crawler-notice evidence
2025-06-25	Kadrey v Meta summary judgment order	Claim-specific evidentiary burden
2025-07-01	Cloudflare Pay Per Crawl and crawler analysis	Crawler logs and access-control evidence
2025-09-05	Anthropic settlement signal	Settlement, claims, deletion, and remediation evidence
2026-03-18	UK Copyright and AI report	Work-level transparency and rights-reservation evidence
2026-05-05	Publisher action against Meta	Institutional rightsholder pressure

Forecast / forward pressure signals

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

Source basis: U.S. Copyright Office Generative AI Training Report; UK Copyright and AI Report; EU GPAI Code and training-summary template; IPTC AI opt-out recommendations; EDRLab TDM Reservation Protocol; Cloudflare and Stack Overflow crawler-control material; public litigation and licensing signals.
Forecast boundary: Qualitative evidence-pressure assessment only.

Pressure	Evidence question	Minimum Evidence Record	Direction	Confidence
Work-level source-lineage demands will intensify	Can a specific work be connected to source, permission, access, training, exclusion, retention, output, and model version?	Work-level source register; Licence and rights-reservation record; Dataset inclusion/exclusion evidence; Model-version linkage	Increase	High
Opt-out and crawler-control disputes will move into log evidence	Can the publisher show what reservation existed, when it existed, which crawler received it, and how the server responded?	Timestamped opt-out snapshot; Crawler logs; Protocol-version record; Access-control response record; Licence/payment event log	Increase	Medium
Licensing disputes will shift toward scope, audit, and model-version evidence	Can the parties show which works were licensed, for which systems, for which uses, and under which audit or exclusion terms?	Licence schedule; Rights-holder authority record; Permitted-use map; Model/version coverage record; Audit trail and revocation record	Increase	Medium

Case studies

Anthropic settlement: resolution creates a new evidence layer

The settlement signal matters less as a liability narrative than as an evidence-operation signal: identifying works, administering claims, recording exclusions or deletion obligations, and defining future-use boundaries become formal records.

Source basis: Reuters Anthropic settlement report; Anthropic settlement administration site.
Evidence boundary: This case study does not determine liability, infringement, settlement fairness, class eligibility, or private dataset contents.

What stronger evidence would have required:

Weak record	Stronger evidence
Settlement headline or aggregate class description	Work-level claim file, identifier list, source status, class-member notice record, exclusion/opt-out log, and settlement-scope boundary.
General deletion or remediation language	Dataset location, copy lineage, deletion/exclusion execution log, model-version impact analysis, and verification record.

EU GPAI summaries: transparency is not work-level verification

The EU transparency model creates public-facing disclosure pressure. The evidential gap is that a summary can tell readers about content categories while still not allowing a rightsholder to verify a named work without underlying records. Summary transparency may satisfy a disclosure layer while still failing the evidential question rightsholders care about: was my specific work used, under what basis, and in which model state?

Source basis: EU GPAI Code of Practice; EU training-content summary template.
Evidence boundary: This case study does not assess compliance by any provider or adequacy of the final regulatory regime.

What stronger evidence would have required:

Weak record	Stronger evidence
Published training-content summary only	Auditable source register, dataset manifest, source-category mapping, licence basis, and exclusion evidence.
Generic copyright compliance statement	Rights-management workflow record, policy-to-source mapping, complaint-handling log, and model-version linkage.

Opt-out evidence: reservation signals need preservation and crawler-specific proof

Opt-out and TDM reservation mechanisms are useful, but the future dispute will often ask when the signal existed, whether the crawler saw it, whether the crawler was bound by it, and whether content was already copied before the reservation was made.

Source basis: IPTC AI opt-out recommendations; EDRLab TDM Reservation Protocol; Cloudflare pay-per-crawl material.
Evidence boundary: This case study does not establish enforceability of any opt-out method or non-compliance by any crawler.

What stronger evidence would have required:

Weak record	Stronger evidence
Current robots.txt or metadata field	Timestamped snapshots, signed policy version, server access logs, crawler identity, HTTP response records, and preservation record.
General do-not-train statement	Machine-readable reservation, licence terms, crawler-specific access policy, and evidence of service or technical exposure.

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	Training, transparency, settlement, and model-linkage disputes	Permission / lineage gap	Work-level source register, model-version linkage, licence and exclusion records
Copyright and publishing	AI training-data and dataset lineage	Publisher actions, opt-outs, licensing, and attribution disputes	Verification gap	Named-work verification route, rights-reservation register, crawler log evidence
Platforms and online services	Platform and third-party record dependency	Crawler monetisation, API content licensing, scraping disputes	Platform-dependent record	Access logs, API terms, crawler identity, response records, independent preservation
Legal and regulatory disputes	IP and commercial conflicts	Fair-use, market-harm, settlement, and discovery pressure	Context gap	Product context, market-use evidence, source acquisition records, claim-to-record map

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
Training source lineage	Work identifier, source URL/system, acquisition date, crawler/API/manual path, dataset membership, transformation record, model-version linkage	Turns broad dataset language into work-level evidence.
Permission and licensing	Rights-holder authority, licence schedule, permitted use, exclusions, duration, territory, model coverage, downstream-use scope, audit rights	Prevents licence announcements becoming unverifiable commercial claims.
Opt-out and rights reservation	Machine-readable reservation, policy text, timestamped snapshot, protocol version, server logs, crawler identity, response evidence	Shows when reservation existed and whether access occurred after notice.
Crawler and access control	Crawler identity, user agent, IP/range, path requested, response code, rate limits, payment/licence status, block/allow decision	Makes platform access and pay-per-crawl claims checkable.
Settlement and remediation	Claim file, work list, class notice record, exclusion/opt-out record, deletion/exclusion logs, model-impact boundary, future-use rule	Shows what a resolution covers and what remains outside it.
Output and attribution	Prompt/output sample, source-comparison method, attribution source, hallucination review, reviewer decision, correction/takedown trail	Separates training-source questions from output, attribution, and reputation questions.

Recommendations by audience

Audience	Practical action
Creators	Preserve original works, publication history, rights-reservation snapshots, opt-out evidence, and licensing communications.
Businesses	Create an AI content-use register covering sources, permissions, licences, exclusions, model/provider dependencies, and logs.
Legal	Map every claim to source, timing, permission, custody, verification, and limits before demands or defence.
Providers	Maintain dataset manifests, ingestion logs, transformation records, complaint workflows, deletion/exclusion records, and model-version linkage.
AI teams	Design source-lineage capture into data pipelines before training, not during discovery.
Public institutions	Require procurement evidence for training-data source, permission, opt-out handling, transparency, and auditability.
Education and research	Preserve dataset provenance, consent/licence status, institutional approvals, and research-use boundaries.
Media and publishing	Maintain rights schedules, crawler policies, opt-out snapshots, licensing negotiations, archive terms, and traffic/crawler evidence.

Methodology and limitations

This report uses selected public sources. It is not global incidence, legal advice, a liability finding, a forensic audit, or cyber/platform telemetry. Private evidence may exist. Absence of public evidence is not proof of absence. Charts are EviWrite classifications. Forecast scores are qualitative unless otherwise stated.

Signals were selected where a public record raised a clear evidence question around source, timing, control, sequence, verification, or proof limits; where the signal was relevant to at least one EviWrite audience category; and where the signal could produce a practical Minimum Evidence Record.

Deep source appendix

The source register in YAML is intentionally structured as an evidential source register, not a bibliography. Each source is classified by source type, source family, module, reliability tier, jurisdiction, evidential relevance, use, non-use, limitations, related claims, and verification notes.

The source spine for this report is: official copyright and AI policy; court and docket records; settlement administration material; public licensing announcements; opt-out and crawler-control sources; and journalism used only as public-signal context.

Source-to-claim map

Claim	Support level	Main support note
AI training-data disputes increasingly require work-level source lineage rather than aggregate dataset language.	Strong	Official policy and docket sources support the pressure pattern; they do not prove private dataset contents.
AI content licensing is becoming evidence infrastructure because scope, exclusions, audit rights, and model coverage must be provable.	Medium-high	Public announcements support visible licensing-market pressure, not confidential contract terms.
AI opt-out and rights-reservation signals require timestamped preservation and crawler-specific records.	Medium-high	Technical sources support evidence design needs, not enforceability or compliance conclusions.
Settlements and remediation promises create new evidential duties around work identification, exclusion, deletion, and future-use boundaries.	Medium	Settlement sources support public claims-process pressure, not liability or full private remediation history.
Crawler logs are becoming commercial evidence as AI access becomes blockable, licensable, or chargeable.	Medium-high	Sources support crawler-control pressure and platform licensing signals, not measured market incidence.
Public AI training-data disputes do not show a single settled legal direction; they show fragmented legal postures, different factual records, different uses, different market-substitution arguments, and different evidential burdens.	Medium-high	Litigation, settlement, policy, and publisher-action sources support fragmented public legal posture; they do not establish final law for all AI training uses.

Assurance and review status

This report 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. Source register, selected-signal register, forecast-signal register, chart limitations, proof limits, and source-to-claim map are included.

Version and audit record

Field	Value
Report number	AITDDM-2026-05
Version	1.0
Period covered	Public signals to 27 May 2026
Prepared by	EviWrite
Status	Draft source-reviewed report source
Report hash	Not issued
PDF hash	Not issued
Source register hash	Not issued
Machine-readable register hash	Not applicable
Receipt	Not issued for this draft/source version