Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study #23

Copyright and training generative models: content transparency and independent auditing

Legal basis for the use of works, licensing, TDM and verifiable liability

Organization: Observatoire ARGA

English version

21 rue de l'Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

October 23, 2026

Contents

Summary for management....5
The main analytical part....5
1. Subject of the study, questions and boundaries of conclusions....5
2. Legal objects: work, performance, database and information....6
3. Lifecycle chain: where exactly do legal risks arise?....7
4. Exclusive rights in EU law: reproduction and communication to the public....7
5. Article 3 of the DSM Directive: Research TDM....8
6. Article 4 of the DSM Directive: General exclusion and right of reservation....8
7. French transposition: article L122-5-3 CPI....9
8. Databases, publisher catalogues and bulk retrieval....9
9. Section 53 of the AI Act: Copyright Compliance Policy....10
10. Section 53 of the AI Act: Public summary of educational content....11
11. The Voluntary GPAI Code of Practice and its Evidence Limits....11
12. Terms, transitional rules and institutional powers....12
13. Machine-readable prohibitions and opt-out technical architecture....12
14. Chain of title: how the right to a specific work is established....13
15. Licensing of training datasets: limits and compensation....13
16. Intermediaries, ready-made datasets, and distributed responsibility....14
17. Training, model memory, and output playback....14
18. Technical verification of memorization: reproducibility and limits....15
19. Content and model removal: what can be promised technically....15
20. Digital Evidence: Preservation, Chain of Custody, and Examination....16
21. USA: Fair use without universal indulgence for education....16
22. United Kingdom: Outcome of the March 2026 consultation....17
23. Singapore and a comparative approach to computational data analysis....18
24. GDPR, personal photos, and double legal verification....18
25. Independent audit method: mandate, criteria and sampling....19
26. Asset Inventory and Data Lineage Graph....19
27. Classification of risks in the corpus: not one category "from the Internet"....20
28. Statistical auditing and honest expression of uncertainty....20
29. Control of the consistency of independent evaluators....21
30. Verifying Supplier Claims: From Marketing to Legal Proof....21
31. Corrections, copyright holder requests and right of reply....22
32. Market Impact Assessment and Authors' Income....22
33. Independence of the Observatoire ARGA: finances and procedural boundaries....23
34. Application to legal analysis, investigation and sanctions....23
35. Recommendations for consultation European Commission Technology and Copyright....24
36. Implementation plan and quality indicators of independent control....24
37. Practical conclusions: what is confirmed and what remains controversial....25
38. Evidence of Market Damage: Causality, Substitution, Alternatives....25
39. Royalties and Collective Distribution: Verifiable Arithmetic....26
40. Accuracy of legal reservation: time and legal effect....26
41. Trade secrets and closed audit: the limits of transparency....27
42. Proof of Origin Format: Compatibility and Portability....27
43. Output Works: Authorship, Derivatives, Licenses....28
44. Application of the AI Act to GPAI modifications: supplier boundary....28
45. Verifiability of contracts without publication of personal and commercial data....29
46. Comparison of disclosed and actually verified indicators....29
47. Open publication and the right to object to an expert opinion....29
48. Double-check of Article 53 of the AI Act and substantive copyright....30

49. Testing multimodal learning: image, sound, video....30
50. Architecture for secure storage of evidence and litigation hold....31
51. Contractual warranties, audit rights and limits of assurances....31
52. Legal dimension of open repositories and software code....32
53. Photo provenance, metadata, and control of false attributions....32
54. Music and Related Rights: Different Layers of Permissions....33
55. Publishing archive, digital libraries and access licenses....33
56. Representativeness of tests: why negative controls are needed....34
57. Decision Points: Stop, Correct, Publish....34
58. Substantive review of EU consultation proposals....34
59. A final model for assessing evidence without false precision....35
60. Specific exceptions: quoting, temporary copying, and TDM....35
61. Collective Rights Management: Repertoire Accessibility and Mandate....36
62. Press publishers' rights and editorial databases....36
63. Institutional and scientific libraries: scientific status is not a general permission....37
64. Reliability of metadata, provenance markers and machine signatures....37
65. Deduplication and Aggregation: The Probative Cost of Losing the Original Trace....37
66. International versions of the work and the right of translation....38
67. Handling corrections and deletions without false technical promises....38
68. The Economics of Compliance for Small Providers and Authors....39
69. Challenging an expert opinion and procedural fairness....39
70. Evidence scale for legal advice and court material....39
71. Consolidated set of minimum measures for the ARGA Observatory....40
Practical applications and research protocols....40
Appendix 1. Passport of the training data corpus....40
Appendix 2. Certificate of acceptance of a new source....40
Appendix 3. Log of rights reservation and opt-out processing....41
Appendix 4. License and Use Matrix....41
Appendix 5. Protocol for proof of legitimate access....42
Appendix 6. Due diligence of the training data provider....42
Appendix 7. Public Reporting Checklist for Article 53(1)(d)....42
Appendix 8. Technical provenance graph and transformation control....43
Appendix 9. Protocol for retraining and updating the model version....43
Appendix 10. Testing the risk of reproduction of protected text....44
Appendix 11. Copyright Claim Registration Form....44
Appendix 12. Chain of Custody for Digital Materials....44
Appendix 13. Statistical sampling plan for a closed case....45
Appendix 14. Checking the agreement of two experts....45
Appendix 15. Procedure for temporary suspension of use of the kit....46
Appendix 16. Verification of corrective actions and residual risk....46
Appendix 17. Register of Independence and Conflicts of Interest....46
Appendix 18. Combined Map of Intellectual Property Rights and the GDPR....47
Appendix 19. Case study: photographic catalog....47
Appendix 20. Case Study: Song Lyrics and a Generative Model....47
Appendix 21. Case Study: Teaching Materials and Translation....48
Appendix 22. Conditional case: software code and open license....48
Appendix 23. Case Study: News and Publishing Archives....48
Appendix 24. Case Study: Museum Collection and Scientific Research....49
Appendix 25. Conditional case: court materials and commercial base....49
Appendix 26. API Provider as a Data Source for Further Training....50
Appendix 27. Model for negotiating a license for a subject catalog....50

Appendix 28. Verification of the powers of a collective organization....50
Appendix 29. Cross-border matrix of legal actions....51
Appendix 30. Audit of claims of “100% legal” training....51
Appendix 31. Compliance Cost Assessment Matrix....51
Appendix 32. Observatoire Public Error Correction Policy....52
Appendix 33. ARGA draft responses for consultation with the European Commission....52
Appendix 34. Secure Data Room Protocol for the Auditor....53
Appendix 35. ARGA Ninety-Day Training Pilot....53
Appendix 36. Template of the final auditor's report....53
Appendix 37. Procedure for requesting clarification from the model developer....54
Appendix 38. Special regime for information on sanctions and international search....54
Appendix 39. Risk checklist before publication of the report....54
Appendix 40. Matrix for monitoring changes after October 2026....55
Appendix 41. Protocol for matching photographic rights with file versions....55
Appendix 42. Test of machine-readable rights reservation when changing URLs....56
Appendix 43. Working example of checking a mixed archive catalog....56
Appendix 44. Comparison of GPAI and Public Summary of Learning Content Versions....56
Appendix 45. Test of author's texts without reproducing the full work....57
Appendix 46. Digital Music Library Agreement Card....57
Appendix 47. Example of license reconstruction through a chain of intermediaries....57
Appendix 48. Instructions for a secure response to a copyright holder's request....58
Appendix 49. Control over the publication of the expert opinion of Observatoire ARGA....58
Appendix 50. Control of the cross-border judicial scenario....58
Appendix 51. Conditional calculation of royalties without arbitrary presumption of damage....59
Appendix 52. Register of versions and critical dates of legal review....59
Appendix 53. Table of TDM and Model Output Delimitation....59
Appendix 54. Matrix of copyright holder objections by content category....60
Appendix 55. Scheme of access of an independent auditor to a closed building....60
Appendix 56. Protocol for stratified sampling of sources....61
Appendix 57. Regulations for checking the version of a document during an EU consultation....61
Appendix 58. Catalog of Prohibitions on the Use of Audit Methodology for Other Purposes....61
Appendix 59. Checking the compatibility of licenses and sublicenses for code....62
Appendix 60. Comparative Opinion Form on US and EU Judicial Practice....62
Appendix 61. Checking the publishing set after changing the license....62
Appendix 62. Protocol for verification of personal data in a licensed archive....63
Appendix 63. Protocol of internal review before approval....63
Appendix 64. Final editorial register of unresolved issues....63
Sources and normative and methodological bases....64

Summary for management

The main finding of the study is that the legality of training a generative model on content is not determined by a single, universal "artificial intelligence" permission. It depends on the verifiable rights to the work and database, the legality of access, the applicability of specific TDM exceptions, the provider's authority, compliance with the reservation of rights, privacy policies, and the model's behavior after release. Neither a public URL, nor a "public source" statement, nor the publication of a summary under Section 53 of the AI Act confirm the legality of each included work. [1][2][3][6]

As of 23 October 2026, Articles 3 and 4 of the DSM Directive apply to the EU, and providers of applicable GPAI models are subject to the obligations of Article 53 of the AI Act, including a copyright compliance policy and a public summary of the training content. The European Commission published a mandatory template for such a summary in July 2025. The voluntary GPAI Code of Practice assists with compliance, but does not in itself constitute judicial proof of non-infringement. Transitional provisions are provided for models posted before 2 August 2025.[1][2][3][4]

Comparative case law requires a precise distinction between actions and procedural stages. In Munich in November 2025, the trial court considered the reproduction of song lyrics in a model and its outputs. In the United States, the June 2025 decisions in Bartz and Kadrey distinguished between the issues of the method of obtaining copies, use in training, and proof of market harm. Moreover, on July 20, 2026, the Bartz v. Anthropic dispute concluded with the approval of a class action settlement and a final judgment on the agreement. The latter does not create a universal solution on the merits for all cases of model training.[11][12][13][22][23]

For Observatoire ARGA, an independent verification is proposed at five levels: a legal map of rights and exceptions; a registry of the provenance of source objects and derivative files; tests of the processing of machine-readable restrictions; tests of the risk of replication in the results; and structured public reporting preserving the boundaries of uncertainty. It is recommended that trade secrets be taken into account through secure access, but not as a substitute for a complete waiver of evidence. The study presents 64 appendices with practical forms, protocols, and case studies. They are not the result of an external audit.

Of particular institutional significance is the Commission's Technology and Copyright consultation, open until 3 November 2026. ARGA's position may focus on technical signals accessible to small copyright holders, transparency in the licensing chain of authority, separate disclosure of public summaries and closed evidence, and low-cost dispute resolution procedures. This report provides an analytical basis for possible participation, but does not in itself warrant the filing of a consultation response. [5]

The final principle	Minimal evidence
Legitimate source	Conditions of access on the date of receipt and legal basis
Respect opt-out	Current technical control and historical logs
Licensing	Object, copyright holder, volume, sublicense and territory
Duties of the GPAI	Rights Policy, Summary, and Model Version Linking
Reliable audit	Reproducible sampling, limitations, independence

The main analytical part

1. Subject of the study, questions and boundaries of conclusions

This paper examines the legal and evidentiary chain of using protected content to develop generative models. The focus is not on an abstract opposition between creativity and technology, but on determining what actions are performed at individual stages: source retrieval, access, copying, normalization, training set creation, training and retraining, storage of intermediate copies, model distribution, and output generation. Each action may involve different exclusive rights, exceptions to them, contractual obligations, and control regimes. A legal opinion on the legality of one stage does not constitute permission for the others. [1][2][6]

The study draws on EU law as of 23 October 2026, the current version of the AI Act, Directive DSM 2019/790, InfoSoc 2001/29, database law, the French transposition, official materials of the European Commission and

limited comparative practice in the UK, US and Singapore. The audit recommendations are a proposal by Observatoire ARGA; they do not constitute a statutory certification and do not confirm the conduct of an audit of a specific company. Court cases are cited by authority, date and extent of their relevance.[1][2][9][10]

The method distinguishes four statuses: mandatory regulation, official clarification or administrative practice, judicial review in a specific case, and research proposal. This method of presentation is especially important in a rapidly changing field, where a voluntary industry code is sometimes mistakenly presented as law, and a first-instance court decision as a definitive, universal rule. Conclusions about compliance with rights can only be drawn from a documented factual chain, and not from the commercial reputation of the developer or the mere fact of publishing a data declaration.

Research question	What needs to be installed	What the evidence does not allow us to conclude
Can the piece be used for TDM?	Legality of access, basis for exclusion, reserved rights status, type of action	What is permitted to publicly reproduce the work in the model response?
Is there a proper license?	Object, copyright holder, territory, methods of use, term, transfer of rights	That the license covers all versions, derivatives and third-party kits
Has Section 53 of the AI Act been implemented?	Copyright Policy and Published Summary of Educational Content	That each individual work was used legally
Can you trust independent audit?	Sample size, access to evidence, limits of verification, and expert independence	That the absence of a violation found proves its impossibility

2. Legal objects: work, performance, database and information

The auditor's first task is to classify not the file itself, but the protected objects it contains. A single web document can contain copyright text, a photograph, an illustration, a database design, a soundtrack, a trademark, and personal information. Interest in "public data" does not eliminate intellectual property rights in the creative form of its presentation. Facts themselves are generally not protected by copyright as a literary work, but the selection, arrangement, and original description may be. A separate sui generis regime for the database producer protects investments in obtaining, verifying, and presenting the content. [6][7]

It's especially important to distinguish between the right to the original photograph and the rights of the individuals whose images are captured in it; the right to the musical composition and to the specific recording; the right of the author of the article and the rights of the press publisher; the right to the software code and the terms of the repository license. The presence of an open URL confirms technical accessibility, but does not confirm either the transfer of rights or consent to any use. An archive access agreement also does not necessarily include the right to extract thousands of materials for the development of a commercial model. The legal map of a dataset must reflect the multiplicity of rights and possible authorized parties.

In debates about generative models, it's useful to distinguish between intermediate copying and the reproduction of essential elements in the final product. For the former, the grounds for access and TDM exceptions are important, while for the latter, the similarity of the protected form, the means of presenting the result, and the possible applicability of other exceptions are important. The absence of the author's name in the search results does not preclude infringement, and the presence of a recognizable theme does not in itself prove it. Expertise should document the elements subject to comparison, rather than assess similarity based on subjective impressions.

Content type	Additional copyright holder or regime	Special control
Book and translation	Author's rights, translator's rights, publishing licenses	Separate versions, languages, scope of permissions
Photography with people	Copyright and possible personal/data rights	Source, consents and publication limits
Lyrics and sound recording	Rights to text, music, performance and phonogram	Separate analysis of text and audio
Database of judicial practice	Database, individual editorial materials	Manufacturer's rights and conditions of access
Software repository	Code ownership, dependency licenses	Compatibility of conditions and preservation of notifications

3. Lifecycle chain: where exactly do legal risks arise?

A typical training chain begins with identifying a URL, directory, or intermediary. This is followed by downloading, unpacking, format conversion, filtering, and deduplication; the material is then distributed among training, evaluation, and tuning. The process may create temporary and persistent copies, embeddings, validation samples, indexes, and logs. An early error in legality is inherited by subsequent derivatives and does not disappear simply because the original file is removed from public storage. In fact, the absence of the original file complicates proving legality. [1][6]

It is recommended to link each transition to a verifiable artifact: source manifest, access date and method, license terms snapshot, scraper version, applicable exemption decision, opt-out status, and a list of products that received the data. This monitoring should also extend to providers of the final datasets. Transferring liability under a contract does not make illegal content legal; a vendor guarantee merely redistributes part of the contractual risk and may prove economically worthless without evidence and insurance coverage.

In a distributed infrastructure, the same entity may appear as a source file, extracted text, a cleaned-up record, and a fragment for further training. Therefore, a standard URL does not serve as a stable identifier. A cryptographic hash of a specific file, a work identifier if available, a parent transformation identifier, source information, and a link to a version of the collection should be stored. This model ensures traceability without requiring publication of a complete list of works and respects commercial confidentiality, but does not allow objects to be hidden from an authorized verifier without equivalent proof.

Stage	Typical risk	Mandatory verification trace
Detection	Mixing accessibility and permission	URL, operator, time, access rules version
Receipt	Violation of restrictions or rights of third parties	Access Basis and Technical Log
Preparation	Loss of attribution and opt-out	Transformation graph, hashes, immutable ledger
Education	Use of unauthorized categories	Version of the set, decision taken, exceptions
Release and update	Lack of information about new sources	Relationship between model version and public summary
Issuance	Reproduction of a substantial part	Safe Test and Response Protocol

4. Exclusive rights in EU law: reproduction and communication to the public

Along with national exceptions, the international "three-step test" of Article 13 of TRIPS must be taken into account: special cases, absence of a conflict with normal exploitation, and absence of unreasonable harm to the legitimate interests of the copyright holder. This does not replace the specific terms of the DSM, but serves as a framework for their interpretation. [19]

Directive 2001/29 establishes a broad right to authorize or prohibit the direct and indirect, temporary and permanent reproduction of a work in whole or in part. For educational purposes, this means that even a technically incomplete copy should not automatically be considered legally neutral. Either the copyright holder's permission or a specific applicable exception, subject to its conditions, is required. The right of communication to the public and making available to the public is assessed separately if the model or service allows the reproduction of protected fragments. [6]

It is necessary to establish where the copying and other actions took place, what national law applies, who controlled the process, and what rights may have been violated. In a cross-border system, a developer may be registered in one country, store data in another, train a model in a third, and provide it to users in multiple countries. This does not boil down to the assertion that "the law of the server country" resolves the entire dispute. The territoriality of exclusive rights, rules on applicable law, and judicial jurisdiction require a separate determination of the facts.

Finally, the copyright holder must prove a relevant protected interest, and the plaintiff must link the disputed use to a specific action by the defendant. The auditor does not replace the court, but can organize the evidence: compare copies, document access conditions, construct a technical chronology, and identify uncertainties. A proper opinion includes both evidence of potential infringement and evidence of permitted use, including licenses, good-faith reservations, or applicable exceptions.

Legal element	Security question	Output limit
Originality	Is there a protected creative form?	Not protecting the facts and ideas themselves
Playback	Has a copy of the protected element been created?	Distinguish between the source, the derived fragment, and the model parameters
Exception	Does the action fall under a specific rule?	Do not transfer the exception to another use case
Message to the public	Is the material reproduced for external users?	Analyze a specific exit and access mechanism

5. Article 3 of the DSM Directive: Research TDM

Article 3 of Directive 2019/790 provides for a mandatory exception for Member States for reproductions and extracts in text and data mining for scientific research purposes by research organisations and cultural heritage institutions. Lawful access is required; copies must be stored with an appropriate level of security and may be retained for scientific research, including verification of results. For this construction, the slogan about the usefulness of the technology is not important, but the status of the subject, the purpose, the lawfulness of access, and the scope of the activity. [1]

A commercial company doesn't become a research organization simply by hiring scientists or publishing a paper. Similarly, having a university partner doesn't in itself provide immunity to the entire technology group. Each project should evaluate the actual roles, control over the results, and the conditions under which the commercial participant gains access to research copies. In mixed initiatives, it is recommended to separate the research environment and the commercial training cycle technically and contractually.

This exception cannot be described as a blanket permission to publicly distribute copies of the corpus or commercialize trained models without additional verification. The research results and the copies themselves are subject to different regimes. A separate log of the copies is required for auditing, ensuring the verifiability of the results while restricting access to individuals not involved in the research project.

Condition	Confirmation	Red flag
Beneficiary status	Charter, actual activities, management	The company is called a university only in advertising.
Scientific purpose	Approved protocol and research questions	Developing a commercial API without research boundaries
Legal access	Licenses, subscription, open source	Downloading from a pirated mirror
Storage	Role model, deadlines and protection	Unlimited access for contractors

6. Article 4 of the DSM Directive: General exclusion and right of reservation

Article 4 permits reproductions and extractions of legally accessible works for TDM purposes other than academic ones. However, the copyright holder has the right to expressly reserve the corresponding use in an appropriate manner. For publicly accessible online content, the article explicitly cites machine-readable media as a typical reservation mechanism. The test of regularity cannot be replaced by the phrase "we downloaded materials that were on Google": legal access and the absence of an effective reservation must be demonstrated in relation to the time and route of use. [1]

It also cannot be claimed that any set of robots.txt rules automatically permits or prohibits all types of learning. The robots.txt file was created primarily to manage access by automated agents; the legal assessment of a specific signal depends on its content, addressee, time, national implementation, and other information about the copyright

holder's wishes. The correct technical approach is to preserve the original machine-readable signal and the result of its crawler interpretation, recording discrepancies and checking access restrictions at the work and domain level.

A reservation may be published after the initial collection, or the work may be available on multiple platforms with different terms. In such cases, a chronology and legal analysis of subsequent transactions is required; neither the automatic retroactive effect of each new opt-out nor the complete absence of any subsequent copying and reuse can be promised. At a minimum, the developer must prevent future unauthorized collection and have a process for handling disputed assets, and the rights holder must have a clear channel for reporting and confirming their authority.

Situation	Procedure	Erroneous conclusion
Machine-readable ban before collection	Record the signal and do not include it in the general collection without legal grounds.	"The site is publicly accessible, so there is no ban."
The signal is inconsistent	Submit for legal review, save the images	"Any robots.txt file completely replaces the law."
Opt-out received after collection	Link to dates, versions, and future actions	"Always retroactive" or "never affects"
Several copyright holders	Check the scope of the applicant's authority	"The editor has all rights to the material."

7. French transposition: article L122-5-3 CPI

In France, Article L122-5-3 of the Code of Intellectual Property distinguishes between scientific TDM and other automated text and data processing methods. Under the general regime, the law permits copies of works with legal access, except in cases of proper objection by the author, particularly in machine-readable form for online content. Copies must be stored securely and destroyed upon completion of the relevant analysis, in accordance with the conditions stipulated by the regulation. For the scientific regime, special rules are provided for the preservation of results for verification purposes. [21]

For the French entity, it is crucial not only to cite the directive but also to compare it with applicable national law and the actual circumstances. The qualifications of the actor, the nature of the work, the category of protected rights, the technical location of the activity, and the applicable license are all relevant. For example, a photobank license may permit editorial use of an image while simultaneously prohibiting its use for training machine learning models. The exception cannot be analyzed in isolation from the method of obtaining the content.

An auditor from the Observatoire ARGA assessing French practices must document the date of copying, proof of access, opt-out content, copies of the terms of use, a note on the legal basis, and the date of deletion of intermediate copies. The absence of this information indicates a lack of evidence, not an automatic violation. This approach allows for reliable conclusions to be drawn in consultations without attributing to ARGA the authority of a court or national supervisory authority.

Action	Criterion under French law	Document
Scientific analysis	The status of the research subject and the scientific purpose	Charter, design documentation, access rules
General TDM	Lawful access and lack of proper objection	Snapshot of conditions and signals log
Saving copies	Distinguish between scientific and general regime	Retention policy and deletion control
Transfer to supplier	Check rights, contract, territory	Licensing register and DPA for personal data

8. Databases, publisher catalogues and bulk retrieval

Even when individual facts lack creative originality, the extraction of large numbers of structured records may affect the sui generis rights of the database producer. Directive 96/9 protects significant investments in obtaining,

verifying, and presenting the contents of the database. The scope of the extraction, the qualitative significance of the selected portion, the frequency of repeated operations, and the effect on the normal operation of the database are analyzed. [7]

The technology team sometimes assumes that copying metadata—case titles, annotations, annotation categories, or library cards—is inherently free. In reality, various elements may simultaneously fall under database protection, editorial annotation copyright, and contractual control over APIs. The distinction between official government sources, editorial services, and commercial aggregators is also important: the right to the original public document does not always cover someone else's commercial systematization.

When evaluating a supplier, you should require a rights map for the dataset itself, not just a certificate of ownership for individual recordings. For each collector, extraction criteria, total volume, speed, frequency, and the presence of an independent protected structure are required. If the same dataset is resold through a chain of intermediaries, both the provision rights and the technical origin must be restored. A commercial transaction with an intermediary without the right to sublicense does not eliminate the defect.

Sign	What's going on?	Legal risk
Bulk catalog upload	Entries, annotations, structure	Significant extraction from the database
Periodic delta crawl	New or modified entries	Repeatability of operations, total volume
Metadata enrichment	Additional editor descriptions	Independent copyright
Resale	Supplier license, sublicense rights	Breaking the chain of rights

9. Section 53 of the AI Act: Copyright Compliance Policy

According to Article 53(1)(c) of Regulation (EU) 2024/1689, providers of general-purpose models must implement policies to comply with EU copyright and related rights law, in particular to identify and respect the reservation of rights in accordance with Article 4(3) of the DSM, using modern technology. This does not imply an automatic right to use any work, but rather an organizational obligation of the provider in addition to existing copyright provisions. This obligation affects the developer depending on its role and status as a provider of the GPAI model on the Union market. [2][8]

An enforceable policy is not limited to a promise to respect intellectual property. It must list sources, access rules, support for machine-readable opt-outs, control over intermediaries, exclusion of illegal sources, the copyright holder's complaint process, a response mechanism for output reproduction, and designated responsible parties. It can be verified through operational evidence: decisions to stop collection, crawler test reports, statistics on disputed requests, and an update log. If the policy exists only as a publicly accessible PDF and is not reflected in the code or contracts, its evidentiary value is minimal.

The obligations under paragraphs 1(c) and 1(d) remain in force for open models: the exception provided for in paragraph 2 of Article 53 for certain open models applies to the documents listed there under paragraphs 1(a) and 1(b), and not to the general obligation to have a copyright policy and a public summary of the training content. A separate framework, including additional obligations, applies for models with systemic risk. A distinction should also be made between the application date of the GPAI requirements for new models and the transition period for previously posted models.[2][3][8]

Policy element being tested	Strong evidence	Not enough
Opt-out processing	URL/asset signal acceptance and exception logs	Unfilled web form
Checking the dataset	List of sources, statuses, selection	"Everything from the open internet" guarantee
Updating models	Compliance with version ID and copyright policy	Policy without version and owner
Complaints from copyright holders	Response time and results for each case	Universal failure without analysis

10. Section 53 of the AI Act: Public summary of educational content

Article 53(1)(d) of the AI Act requires the compilation and publication of a sufficiently detailed summary of the content used for training, using a template provided by the AI Office. The European Commission published the template on 24 July 2025, clarifying its structure: general information about the model, a list of source types, and significant aspects of the processing. This publication serves to enhance transparency and enforceability, but does not automatically require the entire corpus, including every protected work and trade secret, to be made publicly available.[2][3][17]

In its official FAQ, the Commission requires the publication of a summary no later than the placement of the relevant model on the Union market; in future, updates are provided at least every six months, or earlier if the information has changed significantly. For certain models manufactured before 2 August 2025, a transitional period of up to 2 August 2027 applies. These conditions should be linked to the precise legal status of the supplier and version. The public summary may contain justified gaps in relation to older models, but these gaps should not be disguised as complete knowledge of the corpus. [3]

During an audit, it's useful to check the summary against the internal provenance registry: whether the sources of licensed content, publicly available sets, web collection, user data, synthetic material, and changes in shares after fine-tuning are reflected. High precision in figures is impossible if their original assessment is unexplained; false decimal precision requires ranges, a selection formula, and documented uncertainty. The summary itself does not replace proof of legal access and the absence of TDM prohibition at the level of a specific work.

Auditor's question	Public layer	Confidential layer
What types of sources were used?	Categories, main sets, shares/ranges	Asset-level registry, licenses, hashes
Who is responsible for collection?	General diagram of providers and roles	Contracts and contractor control protocols
How is opt-out taken into account?	Description of the procedure and channels	Complete signal and exception logs
What has changed after the update?	Date and nature of the material change	Delta Manifesto and the Decision for a New Cycle

11. The Voluntary GPAI Code of Practice and its Evidence Limits

The Code of Practice for General-Purpose Models contains chapters on transparency and copyright, as well as a separate chapter on security for models with systemic risk. The European Commission describes it as a voluntary mechanism to facilitate compliance with the AI Act. Adherence to the Code or a declaration of compliance is useful for assessing the maturity of procedures, but does not in itself constitute a judicial determination of the legitimacy of the entire model training history. In contrast, the mandatory public summary template follows from the regulatory construction of Article 53(1)(d). [4][17][18]

An independent assessor must distinguish between three issues: whether the supplier has signed the code, whether the promised measures have actually been implemented, and whether specific transactions with copyrighted assets are legal. Failure to comply with a specific measure of the code does not always constitute a direct violation of the AI Act if there is an alternative means of compliance; similarly, a formal signature does not make violations impossible. It is recommended to use a "compliance matrix" linking each voluntary commitment to specific evidence and a separate mandatory provision, if any.

It's especially important to indicate any significant limitations on the auditor's access. If a company only provides a public summary and marketing statements, the report should be called a disclosure review, not a full-fledged origins audit. Independence isn't about a fancy title, but about the ability to obtain a negative result, publish qualifications, and be independent of the audit's commercial outcome.

Statement	What really confirms	What does not confirm
"GPAI Code Signed"	Voluntary commitments have been made	Legality of each training file
"The summary has been published"	The requirement for public disclosure to an appropriate extent has been met	The completeness and authenticity of the hidden registry
"The auditor found no violations."	No violations were found in the specified sample	Absolute absence of violations

Statement	What really confirms	What does not confirm
The corpus has an open license.	The presence of some licensing conditions	That all objects are free from other restrictions

12. Terms, transitional rules and institutional powers

For GPAI providers, the obligations under Article 53 began to apply on 2 August 2025, and the Commission's powers to ensure them began on 2 August 2026; for models placed on the market before the first date, specific transitional periods are set out, as reflected in Article 111 and the Commission's clarifications. The question of the specific version of the model and the moment of its placement on the market is mandatory: the same deadline cannot be automatically applied to all products, or subsequent additional training always considered legally irrelevant.[2][3][8]

The AI Office oversees the relevant obligations of GPAI providers under the AI Act, while copyright disputes, licensing, compensation for damages, and specific remedies are handled under applicable copyright, procedural, and national law. The power to request documentation under the AI Act does not automatically grant a civil plaintiff access to any closed dataset in its entirety. Conversely, the publication of a public summary does not preclude the right of a court or other competent authority to request information where there is a procedural basis.

For internal control, it is recommended to maintain separate calendars: one for the AI Act's regulatory requirements, a second for license deadlines, denials, and legal claims, and a third for technical review cycles of the corpus. Mixing these calendars leads to errors, such as replacing the obligation to update the summary with a statement about "re-obtaining consent from each author every six months." This obligation, as stated, is not reflected in the Commission's template.

Control date	Meaning	Control action
August 2, 2025	Commencement of application of GPAI duties for relevant new models	Checking model status and version
August 2, 2026	Powers of the Commission to ensure compliance with GPAI requirements	Check the readiness of evidence and procedures
August 2, 2027	Transitional benchmark for previously posted GPAI models	Document applicability and gaps
Every six months or earlier	Update of the public summary with new training data on the Commission's clarification	Match the model version with the delta manifest

13. Machine-readable prohibitions and opt-out technical architecture

A rights reservation must be communicated to the technical system not as an unexplained check mark, but as an event with a source, time, volume, and degree of trust. Signals at the level of a site, URL, collection, specific file, and standardized metadata are possible. The technical format must be assessed in the legal context of Article 4(3) of the DSM and the French implementation, rather than assuming that there is a single legally sufficient signal for all types of content. In particular, search indexing signals are not automatically identical to signals addressed to model training. [1][21]

The process must record the robot's request, the response received, the hash of the rules snapshot, the rules for handling redirects and rechecks, and the reason for approval or blocking. Mirror sites, CDNs, archived copies, and republications pose a separate challenge: the copyright holder can opt-out of the original source, but a copy of the work continues to circulate through an intermediary without the same signal. Therefore, the registry must maintain not only the URL but also the identity of the object, including visual and textual fingerprints, with caution against false matches.

A verifiable minimum standard is proposed for public consultation: a publicly accessible signal specification, the ability to object without a paid intermediary, collision detection and resolution logs, and tests for compliance with restrictions during reassembly. At the same time, the standardization must be accessible to small publishers and

authors without technical teams. A system in which technical complexity effectively deprives copyright holders of their right to choose cannot be built.

Signal channel	Technical risk	Auditor control
Site rules	The robot is using an outdated cache.	Timestamp, TTL, recheck
File metadata	They are lost during conversion.	Checking before and after conversion
Register of works	Copyright Holder Linking Error	Authority verification procedure
Notification directly	Incomplete coverage of derived sets	Asset lineage and routing confirmation

14. Chain of title: how the right to a specific work is established

A corpus provider can guarantee that all materials are their own creation or legally licensed, but this contractual assertion is no substitute for verifying the chain of rights. For a work of art, the author and any potential assignment or licensing agreements must be established; for a music track, the rights to the composition and recording; for a book, the relationships between the author, translator, publisher, and digital distributor. It is important to distinguish between the right to release content to users and the right to authorize its copying for machine learning.[6][7]

The audit begins with selecting the level of granularity. For legal analysis, it is desirable to have the work's identifier, specific version, rights category, title holder, and licensee. If there is no legal chain, the uncertainty should be flagged, and the asset should be removed from high-risk streams until confirmation. Random testing should not be limited to large, well-known publishers: smaller authors and archival materials often have more complex or less well-digitized contracts.

A particular problem is licenses with the right to sublicense. The provider may only have the right to display the work in its own interface and not the authority to transfer the asset to another developer. Recommended verification: request a certified description of rights, review a selection of source agreements, and verify exceptions and restrictions by territory, term, type of use, and model. Doubts about rights are not resolved by the provider's high commercial rating.

Chain element	Confirmation	Consequence of the gap
Authorship and original rights	Agreements, registries, creation metadata	Uncertainty of asset status
Transfer/license	Object and specific uses	Inability to prove eligibility for education
Sublicense	Direct authority to transfer rights	Risk of illegal downstream use
Territory and term	Jurisdictions and dates	Discrepancy with the geography of training and placement
Right to updates	Conditions for new versions of the work	Inadmissibility of automatic license succession

15. Licensing of training datasets: limits and compensation

A commercial license provides a more specific basis for use, with the correct scope and authority of the parties. However, the word "license" itself does not specify the permitted methods of processing: copying, preprocessing, training, fine-tuning, storage of intermediate copies, transfer to a contractor, use for evaluation, and possible retraining must be explicitly assessed. Some licenses require attribution, limit model categories, or prohibit use in products that compete with the copyright holder. These restrictions cannot be ignored when creating mixed corpora.

For negotiations, a registry of permissions should be created at the asset level, rather than a single line item stating "corpus licensed." Payment may depend on volume, term, number of products, territory, or confirmed use cases. An economic comparison of different remuneration models is only possible if it is clear who pays administrative costs, how payments are distributed among co-authors, and what happens to previously used content after termination.

Observatoire ARGA proposes separately assessing the availability of compensation for small authors. Collective licensing reduces transaction costs, but requires clarity regarding repertoire, representative powers, withdrawal

rights, and distribution of funds. No collective management model should be declared mandatory for everyone without a legislative basis. A practical recommendation is to preserve competition among voluntary licenses and the author's evidentiary right to verify inclusion in a protected set, without requiring public disclosure of the entire closed corpus. [5][9]

License conditions	Question about the contract	Checking execution
Copy volume	Are technical and durable copies allowed?	Register of operations and derivative forms
Training and assessment	What models, stages and tasks are covered?	Manifestation of Versions
Sublicensing	Is the contractor/partner allowed?	List of downstream entities
Reward	Calculation model and control data	Audit formula, reporting
Termination	What to do with copies and existing scales?	Transition plan and feasibility assessment

16. Intermediaries, ready-made datasets, and distributed responsibility

The learning chain often involves independent collectors, cloud storage providers, dataset sales platforms, model developers, and application providers. It's important not to apply a single characteristic to everyone: one participant determines the sources and legal basis, another merely stores copies, and a third modifies the model for a new market. Liability depends on the specific action, contract, and applicable law, not on the chosen role's trade name. [2][8]

The auditor must construct a supply chain, identifying decision points: who authorized collection, who accepted the opt-out, who had the opportunity to delete the disputed material, who initiated the training, and who publishes the summary. Often, the supplier claims to have a "clean" dataset, relying on their own standard without disclosing the criteria. A reasonable audit should include a representative sample of primary sources, licensing documents, and the ability to challenge misclassification.

Contracts should distinguish between the obligation to provide evidence and the obligation to indemnify. Even a broad obligation to indemnify claims is of no help if the provenance log or original metadata are missing. A commitment to cooperate with regulatory review and document preservation is often more important than an abstract declaration of absolute legal purity. A plan for what to do if the supplier goes out of business but the model continues to be distributed is also important.

Subject	Possible control	The requested evidence
Collector	Sources, rules of evasion, moment of copying	Crawl logs, screenshots of conditions
Aggregator	Assembly and resale	List of upstream licenses
Model developer	Training and release	Data manifest and model version
Integrator	Fine-tuning, RAG, API	Additional sets and permission limits
External auditor	Verification method	Selection, limitations, conflicts of interest

17. Training, model memory, and output playback

A model can reproduce a significant protected fragment, despite its architecture not being a standard sequential file database. Technically, this is related to memorization and regurgitation, but the legal qualification depends on the specific law and evidence. In its ruling of November 11, 2025, in the case of GEMA v. OpenAI Group companies, the court found, under the circumstances examined, that parts of song lyrics were reproducible in the model and its responses and upheld a substantial portion of the claims. The court specifically noted that the decision had not entered into legal force as of the date of the announcement. This result cannot be unconditionally transferred to any models or other genres. [11]

Verification requires controlled test queries, recording of the model version and generation parameters, comparison of the output with the documented original, and the exclusion of accidental similarity to a widely

known, unprotected formula. A distinction must be made between a simple thematic response, a retelling of facts, stylistic similarity, and an extended, close reproduction of the protected form. Simply indicating a percentage match without describing the basis for comparison is not suitable for a legal opinion.

The role of the provider and user in creating a disputed response also cannot be resolved by a single rule. The nature of the requests, security settings, predictability of the response, and the provider's ability to prevent replay are all important. An independent procedure requires maintaining a test log without publicly publishing long protected fragments, which could themselves become the source of a new violation.

Result type	Testing method	Possible conclusion
Factual retelling	Verification of a protected form	Generally not sufficient to conclude copying
A short typical phrase	Analysis of originality and prevalence	Low evidential strength of coincidence
A long unique passage	Letter-by-letter comparison, model version	Basis for a subject-matter legal assessment
Playback on simple request	Multiple independent repetitions and settings	Increased risk predictability

18. Technical verification of memorization: reproducibility and limits

A memorization audit cannot be conducted with a single successful or unsuccessful prompt. To ensure reliable results, a sample of works must be selected before testing, separating common, rare, and high-risk fragments, and the number of attempts and query strategy must be recorded in advance. Otherwise, the expert may inadvertently select prompts that produce the desired result and create the impression of high or zero recall. The report separately indicates the frequency of exact fragments, the significance of the match, and the variability of the response.

For text models, normalized sequence matching is useful, but automatic metrics are no substitute for originality analysis. For images, criteria of compositional similarity and an assessment of characteristic protected features are required, while for audio, separate experts on the recording and musical composition are required. Tests in which the examiner inserts the entire protected original into the prompt and then considers its appearance in the output as evidence that the model learned it during training should be excluded.

Results should be presented with confidence intervals and caveats. If no close replication is found in 100 independent trials, the statistical upper bound for the probability of an event under the simplified binomial model is approximately 3% at the 95% level (the rule of three), not zero. This is a training example, not the actual measured frequency of a specific model; the independence of trials and sample quality require separate confirmation.

Protocol Component	Rule	Why is it important?
Sample	Lock before launch	Prevents cherry-picking
Prompts	Separate simple and targeted attacks	Distinguishes between normal and intentionally induced behavior
Parameters	Version, temperature, context, date	Ensures reproducibility
Metrics	Precise fragments, qualitative significance	Avoids false conclusions from the one percent
Conclusion	Risk and limits, not absolute guarantee	Does not turn statistics into a legal presumption

19. Content and model removal: what can be promised technically

After discovering an unauthorized work, the developer can remove it from future training datasets and indexes, block specific output fragments, exclude accessible copies, and begin machine unlearning research. However, it's impossible to guarantee that all influence of the work has been removed from previously trained weights without testing. The effectiveness of removal depends on the architecture, the nature of the data, the stage of training, and

the method used. Therefore, a draft contractual clause stating "complete removal from model memory within 24 hours" may promise a technically unachievable result.

A practical policy should distinguish between categories: source files, intermediate copies, future run datasets, RAG databases, caches, published weights, production logs, and derivative models created by third parties. For each category, specify the actual deletion capability, confirmation, and necessary mitigation measures to mitigate residual risk. Deletion of a copyright holder request record should not be confused with deletion of the data itself.

In disputed cases, suspension of use, additional training, filtering of output, and termination of version support are permitted. The choice of action should be proportionate to the risk, evidence, and available authority, not commercial convenience. The external auditor must receive a test report after the action and the right to indicate that the residual uncertainty has not been resolved. [11][12]

Object	Relative controllability	Mandatory proof of measure
Source file	Typically high when monitoring storage	Deletion and backup logs
RAG index	Typically high with controlled re-index	Version index before/after
Training set for the next cycle	High with a working asset registry	Manifest exclusion
Parameters of the already released model	Not guaranteed	Results of special tests and disclaimer
Open copies of third parties	Limited	Notifications to partners and assessment of residual risk

20. Digital Evidence: Preservation, Chain of Custody, and Examination

In a copyright dispute over AI data, it's especially easy to destroy important evidence by patching the system. Therefore, after a plausible claim, it's necessary to legally preserve relevant access logs, corpus version information, original license terms, crawler configurations, test query sets, and output versions. This doesn't mean permanent storage of all personal data: evidentiary obligations should be coordinated with data protection and the legal retention regime. [20]

The independent expert records the time in the agreed-upon time zone, the file's origin, the cryptographic hash, the person performing the action, the method of acquisition, the immutable audit trail, and access restrictions. If a screenshot of a page was taken after its content was modified, it must be indicated that this is a subsequent state, not an actual snapshot as of the training date. A copy of the repository received from the applicant cannot be used as proof that this version was contained in the original set.

In judicial and regulatory proceedings, it is necessary to distinguish between expert testimony, technical expertise commissioned by a competent authority, and internal corporate investigations. ARGA may prepare independent analytical materials within the limits of its competence and legal access, but it shall not grant the procedural status of a court expert or claim full access to closed model scales if such access does not exist. Evidentiary value depends on disclosed methods and reproducibility, not on the length of the document.

Proof	Fixed fields	Possible distortion
Screenshot of the site rules	URL, date, browser, archive	Removed after controversial collection
Crawl magazine	Agent, IP, timestamp, object	Incomplete storage coverage
Manifest Corpus	Hash, version, dataset ID	Loss of connection during transformations
Test output	Prompt, version, settings, log	Post-observation sample substitution
License	Full text and appendices	Unknown sublicense restrictions

21. USA: Fair use without universal indulgence for education

In the United States, fair use analysis requires consideration of the specific use, the nature of the work, the volume and substantiality of the copying, and the impact on the actual or potential market. In its June 23, 2025, ruling in Bartz v. Anthropic, a federal court distinguished three types of activity: the use of books for instructional purposes, the internal digitization of lawfully acquired print copies, and the creation of a permanent library of pirated copies. The court found fair use for the first two categories on the facts presented, but did not extend such

protection to the permanent pirate library. This was an interlocutory ruling on specific issues, not a general instructional license. [12] On July 20, 2026, a different judge in the same federal court granted final approval of the class action settlement in this case, entered judgment on its terms, and dismissed the case without allowing any further litigation of the claims covered by the settlement. This is a procedural conclusion by agreement, not a reexamination of the universal fair use test or a ruling on the legality of any model. Therefore, when quoting Bartz, it is necessary to distinguish between the June 2025 determination and the final approval of the agreement in July 2026.[22]

In Kadrey v. Meta, the court decided cross-motions for partial summary judgment on June 25, 2025, applying fair use to certain copying within the evidence presented by the parties. The court found insufficient evidence of the alleged market harm in this procedural context, particularly potential market substitution or dilution. This definition cannot be cited as a final immunity from all claims related to training, or as a universal rule for other corpora and markets. Individual issues and claims may have different procedural fates. [13][23] The 2025 US Copyright Office report provides analytical assessment and recommendations, but does not constitute law, a judicial decision, or a blanket authorization for training without a legal basis test. [10]

For the transnational model, this means it's impossible to apply the American fair use doctrine as a direct replacement for the European TDM exceptions. The practical matrix must separately examine the country of application, applicable law, method of obtaining the content, the copyright holder's market, and the existence of a valid licensing mechanism. A commercial team's decision that "training can be done in the US, therefore it's legal in the EU" is legally insufficient and requires consultation with specialized lawyers.

American material	Established narrow meaning	Inadmissible extrapolation
Bartz v. Anthropic (2025)	Different qualifications of training and pirate library	"Pirate sources don't matter"
Kadrey v. Meta (2025)	Materiality of proven market harm	"Any commercial training is fair use"
US Copyright Office Part 3	Expert analysis of politics and law	"The New Mandatory Law of the United States"
EU DSM Articles 3/4	Special European conditions for TDM	"The full equivalent of American fair use"

22. United Kingdom: Outcome of the March 2026 consultation

On March 18, 2026, the UK government published a report on copyright and artificial intelligence, as required by national legislation. It found that the initially preferred idea of a broad opt-out exclusionary rule was no longer the government's preferred course, following strong resistance, primarily from the creative industries. Instead of announcing the enacted law, the government cited the need for further examination of evidence, licensing markets, and technological transparency. [9]

Therefore, as of the date of this study, the discussed UK opt-out cannot be presented as an already established universal exception for commercial education. Current national law, treaties, and existing exceptions should be analyzed, with special attention to potential legislative proposals and consultations. The UK report is also useful from a methodological perspective: it describes economic uncertainties, the difficulties faced by small content creators, and the potential role of transparency in protecting rights.

For ARGA, the important conclusion is that the effectiveness of rights reservation cannot be assessed without measuring the actual share of works for which the author is able to express his/her will and the percentage of technical compliance with the signals. A legal regime that benefits the largest catalogue holders may leave small creators with unaffordable costs. This risk should be factored into the assessment of future EU policy and international consultation positions.

Question	Confirmed status in March 2026	What needs to be studied further
Wide opt-out exception	No longer the preferred government option	Effect on authors and developers
Input data transparency	Directions for further work	Level of detail and cost of compliance

Question	Confirmed status in March 2026	What needs to be studied further
Licensing	Existing and developing market	Accessibility for small participants
Territory of operation	National analysis separate from the EU	Cross-border content flows

23. Singapore and a comparative approach to computational data analysis

Singapore's Copyright Act 2021 specifically regulates computational data analysis: Section 243 defines analysis as identifying and extracting information or using works as examples to improve a computer program, while Section 244 sets out conditions for acceptable copying. This should not be interpreted as an unconditional freedom to download any work from any source: the law has requirements that must be verified based on the text and facts. [15]

Comparative analysis is only appropriate for identical legal actions. For example, the permissibility of creating analytical copies does not automatically imply the right to distribute original works or to ignore confidentiality obligations. When using a foreign dataset, the developer must determine the jurisdiction in which the copies were created, where the subsequent training took place, and what rights are claimed in the target country's market. Differences in national exceptions can create a conflict between the legality of upstream processing and downstream distribution of the model.

ARGA's international expert position emphasizes avoiding "country-hopping" as a synonym for legal compliance. A more sustainable practice is a unified map of origin and permits, supplemented by specific national qualifications. This ensures the ability to pass an audit even when jurisdiction changes and reduces the cost of re-auditing the entire chain.

Comparison criterion	EU	Singapore
Basic design of TDM	Exceptions to Articles 3 and 4 of the DSM	Computational data analysis, Copyright Act 2021
Public opt-out	Significant for the Article 4 regime	An analysis of individual conditions of national law is required
Proof of access	Lawful access	Check the prerequisites provided by national law
Cross-border	Take into account the rights of other states	Do not equate a domestic exception with a worldwide license

24. GDPR, personal photos, and double legal verification

A training dataset may simultaneously contain copyrighted works and personal data. Successful verification of the TDM does not negate the obligation to comply with the GDPR, if it applies. For example, a journalistic photograph may be licensed by a publisher for a specific purpose but contain identifiable individuals; a court case record may be publicly available under a national publication rule but include information about criminal prosecutions and other sensitive circumstances. In such cases, independent processing grounds, transparency, accuracy, minimization, and applicable special restrictions are required. [20][16]

The risk is particularly high when training generative legal assistants using court archives, investigations, and publications on suspected crimes. The copyright license for the information does not automatically permit the reprocessing of personal information about the suspect. For individuals who are not public figures, the global use of information could lead to disproportionate secondary distribution. A separate legal registry for IP and data protection, linked by a single asset ID, is needed.

A technical auditor should verify the purpose of the array and the nature of the data before loading it into an external model. Even if the established license is sufficient for copyright, transferring texts to a foreign API may raise separate requirements for personal data, privacy, confidentiality, and international transfer. The report should contain two distinct results: an assessment of intellectual property rights and an assessment of the permissibility of processing personal data, without reducing one to the other. [16][20]

Situation	Copyright issue	Data protection issue
Licensed face photography	Is training allowed?	Is it legal to process a person's image?

Situation	Copyright issue	Data protection issue
Court material	Editorial text/database rights	Article 10 of the GDPR and publicity of proceedings
Client's letter	Copyright of the author of the text	Confidentiality and legal basis for processing
Publishing on the open web	TDM and opt-out	The principles of the GDPR remain applicable

25. Independent audit method: mandate, criteria and sampling

An independent content origin audit begins not with model test requests, but with a mandate and criteria agreement. The mandate must specify legal entities, models and versions, the training period, geography, source categories, applicable regulations, the scope of access to classified information, and the auditor's right to issue a negative opinion. If the client pre-limits the scope of the audit to a public summary, this is acceptable as a desk review, but it is unacceptable to refer to it as a full licensing audit or asset-level provenance. [2][3]

The methodology should be risk-based. A complete review of every element of a multi-billion-dollar corpus is often technically impossible; the sample should allow for an assessment of the documented risks without concealing the likelihood of missing data. For the formation of strata, it is recommended to use the type of right, provider, collection time period, legal basis, language, content with high commercial value, claims history, and the presence of opt-out signals. Elements selected after a violation has been detected should be identified as targeted, rather than mixed with random sampling.

The report must separately show the sample design, the observed frequency of defects, unresolved gaps, and legal qualification. The percentage of objects without a full license certificate is an indicator of a documentary gap, not an automatic percentage of illegally used works. However, the lack of evidence of legality precludes a positive conclusion of full compliance. The recommended scale is: confirmed, conditionally confirmed, incomplete information, identified noncompliance, and beyond the scope of access.

Program element	Minimal solution	The Error That Prevents
Borders	Models, versions, sets, terms, standards	A false sense of being audited by the entire company
Selection	Random and target strata separately	Hidden cherry-picking
Access	Public and closed sources of evidence	Identification of summary with registry
Grade	Law, fact, technical uncertainty	Substitution of a judicial opinion with an auditor's opinion
Publication	Degree of confidence and limitations	A "certificate" without clear content

26. Asset Inventory and Data Lineage Graph

The recommended unit of observation is not the URL per se, but rather the "object-version-rights-source" unit. Each asset should have an internal identifier, content type, hash, source and acquisition date, license/exception information, legal signal history, relationships to the original and transformed versions, and a list of models in which it was used. These fields allow us to reconstruct the path of a disputed work from a website to a specific training run. Without a provenance graph, removal and verification of downstream usage become a manual investigation based on employee memory.

The architecture must distinguish between legal and technical identifiers. An ISBN identifies a publication, but not always a specific digital file; URLs change; a visual fingerprint can erroneously combine different photographs; a hash captures precise bytes but does not recognize format changes. Therefore, a combination of identifiers and a transformation log are recommended for significant objects. Each deduplication, tokenization, OCR, or slicing operation must maintain a link to the parent object and any applicable constraints.

To be GDPR-compliant, the registry must minimize personal information, especially when dealing with copyright claims and legal cases. The full text of the protected work need not be duplicated in the audit report: a control

hash, secure proof of existence, and protocol-based access by the auditor are sufficient. An independent auditor has the right to report a lack of access, but must not disclose commercially sensitive directories without justification. [16][20]

Field	Content	Integrity check
asset_id/version	Persistent ID and file version	No collisions or gaps
parent_id	Sources of transformation	Backtracking capability
legal_basis	License/Exemption and Status	Correspondence to action and time
reservation	Source and version opt-out	Snapshot of the state on the date of collection
model_lineage	Version/launch usage	Independent coordination with the teaching manifesto

27. Classification of risks in the corpus: not one category "from the Internet"

The corpus should be divided according to verifiable criteria: licensed directories, open-source authors with clear terms, data obtained through TDM exceptions, aggregator collections, and assets of unclear origin. Pirated repositories, illegal mirrors, materials from closed communities, and sources with technical limitations are marked separately. This classification is needed not for an impressive dashboard, but for determining legal and technical oversight.

Risk is determined not only by the probability of unlicensed copyright but also by potential impacts: exceptional economic value, difficulty in compensating the copyright holder, presence of personal information, likelihood of long-term reproduction, and distribution via multiple downstream models. A conditional risk assessment can be represented by a qualitative matrix if statistical estimates lack a proper empirical basis. It is impossible to derive "exactly 82 points of legal security" from arbitrary weights of attributes.

High risk should be mitigated by documentary verification, exclusion of the disputed source, a special license, or cessation of use. It's useful to consider circumvention mechanisms during the audit: a crawler may block the primary domain but continue to access a copy in a publicly accessible archive. Therefore, the assessment should consider the work's persistent identity and republication routes, not just adherence to a specific list of domains.

Category	Recommended status	Necessary procedure
Confirmed license	Tolerance within the scope	Checking the date and version
Legitimate TDM without reservation	Conditional admission	Proof of access and signals
Unknown chain of rights	Quarantine	Further examination or exclusion
Obviously a pirated source	Prohibition of use	Blocking and investigating previously created copies
Established claim	Individual solution	Preservation of evidence and evaluation of measures

28. Statistical auditing and honest expression of uncertainty

For statistical inference, the population should be defined in advance: all records in the corpus, unique works, documents from a single provider, or specific versions of works. These values are not interchangeable. For example, a single literary text found 500 times online may represent 500 files, but only one protected work. To assess legal defects, it is recommended to report metrics for both files and unique assets, if both identifiers are available.

Let's assume that an independent random sample of 200 objects found 12 objects with missing evidence of legitimate access: the observed proportion is 6%. This does not mean that 6% of the corpus contains violations. First, the category denotes a documentary gap; second, the statistical uncertainty interval is nonzero; third, the results are unrepresentative if the randomness of the selection was violated or a significant portion of the corpus was unavailable. A separate indicator of refusal to provide information should be reported.

For complex sampling designs, stratified weights and clustering checks are used. For rare, severe violations, targeted testing may be more efficient than random search, but its results cannot be extrapolated to the entire

corpus using the usual proportions. The report distinguishes systemic procedural deficiencies from episodic defects. This ensures a fair threshold for conclusions and allows for comparison of audits over time without artificially inflating success. [10]

Case study	Calculation	Acceptable output
12 out of 200 do not have access confirmation	12/200 = 6%	The proportion of gaps in the sample that are not violations in the entire corpus
0 out of 100 plays	Rule of three: ~3% upper limit	Zero outcome does not equal zero risk
8 out of 80 targeting tests	10% of the selected risk objects	You can't transfer 10% to a random population
40% of the corpus is not available to the auditor	Coverage 60%	The positive conclusion is limited to the known part

29. Control of the consistency of independent evaluators

Legal classification of licenses and technical classification of matches require human judgment. If two auditors independently review the same asset and one finds copying permitted under the license, while the other points to its restricted subject matter, the system should not conceal the disagreement but document its basis and submit it for legal review. A simple match rate is insufficient when almost all objects fall into an obvious category.

The practical protocol includes assessor briefings, general criteria, blind assessment of a pre-selected control sample, agreement calculations, and a review of discrepancies with the final category being determined. Cohen's kappa coefficient may be used for nominal categories, but only with correct assumptions and a transparent labeling matrix. High agreement does not guarantee legal validity if all assessors used the same erroneous instructions.

A separate role should be responsible for updating the legal section whenever the law changes or a court decision is received. Retroactively changing historical records without preserving the previous version is prohibited. Otherwise, subsequent "high consistency" may reflect a retouched history rather than a mature expert procedure.

Control	Method	Documented result
Blind markings	Two evaluators without access to a colleague's response	Initial categories
Coherence	Intersection matrix and relevance coefficient	Quantitative and qualitative assessment
Disputes	Reasoned legal review	Agreed decision and disagreement
Changing the norm	Method version and date	Historical comparability

30. Verifying Supplier Claims: From Marketing to Legal Proof

Common statements like "we don't use data without permission," "our corpus is secure," or "we only use open-source training" don't carry the same evidentiary value. An audit must transform them into verifiable assertions: are there sources with disputed access, how were opt-outs verified, from which providers were licenses requested, what exceptions were applied, and who approved the disputed categories. Simply publishing a policy without an implementation report does not prove compliance. [2][3]

It is recommended to request specific artifacts for each claim: a signed contract, a legal exclusion matrix, a crawler denial, a training manifest, and a complaint response protocol. If evidence is not disclosed due to commercial confidentiality, a confidential clean room procedure and publication of an aggregated conclusion are possible. However, verbal assurance from a company representative cannot replace an independent review. The auditor must indicate which evidence was not available and what conclusions cannot be drawn from this.

Conflicts of interest when funding an audit by a supplier should be assessed separately. If the auditor's remuneration is contingent on a positive audit report, the credibility of the audit is undermined. The procedure should include a fixed fee for the work, the right to publish a negative report, documentation of the consulting relationship, and verification of the financial dependencies of employees.

Claim	The proper question	Example of proof
All rights reserved	In what area and according to what standards?	Legal-basis asset sample
"A transparent summary has been published"	What versions does it cover?	Public summary + manifests
Opt-out is working	Has the technical implementation been verified?	Negative tests crawler
"Independent audit"	Who pays, approves and publishes?	Engagement letter and conflicts register

31. Corrections, copyright holder requests and right of reply

A proper system must have a procedure for accepting reasonable requests from authors, heirs, publishers, and collective management organizations. An ordinary photographer should not be required to provide a closed copy of a model's entire body weight in order to report a possible infringement. A channel allowing the identification of the work and the claimant's interest, while adhering to the principle of data minimization, is sufficient. A company has the right to verify the authority of the representative and the existence of valid rights, but this verification should not become an endless hurdle.[5][9]

Upon receipt of a request, the date and subject matter must be recorded, relevant evidence must be temporarily preserved, the source lineage and grounds for use must be verified, a reasoned response must be provided, and the proposed action must be documented. Possible responses include refusal with an explanation of the legal basis, limitation of future collection, licensing, de-indexing, contractual settlement, or appeal to a competent authority. The response must distinguish between the confirmed presence of the work in the collection and a presumption based solely on similarity of the source text.

The Independent Observatoire can assess the overall accessibility of the procedure and the quality of the response, but it does not promise the copyright holder compulsory compensation or declare the company's guilt in lieu of a court hearing. The copyright holder must retain the right to challenge an insufficiently reasoned decision, and the company must be able to correct an erroneous public statement. Publication of the names of the parties is permitted only on a specific legal basis.

Stage	Quality control	Output document
Receipt	Confirmation, number, subject	Appeal card
Verification	Authority, object identity	Verification protocol
Study	Lineage, conditions, playback test	Reasoned assessment
Solution	Measure and limits of execution	Reply to the applicant
Appeal	Independent review	Registry of corrections

32. Market Impact Assessment and Authors' Income

One of the central questions in the Commission's consultation is whether AI training can replace demand for original works and how fair compensation might work. There is no reliable universal ratio of "one used file equals a certain amount of lost income." The markets for photography, music catalogs, translation, specialized news, data licensing, and software code vary significantly. The analysis must take into account the works, the quality of substitutes, the structure of contracts, and the counterfactual development of demand. [5][9][10]

It is recommended to measure not only the volume of training data but also observable licensing conditions: the availability of licenses for small authors, tariff dynamics, exclusivity requirements, the presence of intermediaries, legal defense costs, and the role of collective management. Moreover, the data from the parties themselves is potentially relevant and requires independent verification. Complaint statistics may increase due to improved access to protection channels, rather than an increase in actual violations.

The ARGA research proposal recommends several scenarios: a basic market without a new mechanism, a voluntary licensing market with minimal transparency, enhanced documentation of training corpora, and a standardized opt-out with effective controls. Each scenario should be accompanied by assumptions and uncertainty intervals, avoiding unfounded projections of multibillion-dollar compensation.

Indicator	What characterizes	Possible trap
Revenue from licenses	Completed transactions	Does not show all missed opportunities
Catalog price	Market conditions	Distorted by exclusivity and rights package
Complaints about training	Availability and activity of protection	Not equal to the established violations
Output matches	Technical risk of reproduction	Damage to the market is not automatically measured

33. Independence of the Observatoire ARGA: finances and procedural boundaries

Observatoire ARGA may conduct expert research on copyright and AI as part of its analytical program, but it cannot present itself as an accredited certification body, government regulator, or forensic expert unless its credentials are confirmed. It is especially important to distinguish between public monitoring and paid consulting projects: if the company being audited simultaneously pays for other services to the organization or its partners, the conflict of interest must be registered and assessed by an independent entity.

The proposed model includes a scientific supervisor, a legal reviewer, a technical expert, an evidence secretary, and a conflicts of interest committee. The commercial manager must not be allowed to single-handedly edit the conclusion in favor of the paying party. Access rights, rules for publishing the methodology, inconsistencies, and material reservations are defined for each project. If a public report is not possible due to confidentiality, the published statement must clearly indicate the depth of the review and the nature of the classified information.

A key criterion of independence is the ability to honestly reach a negative conclusion without sanctioning the researcher. The report must not imply that ARGA has direct access to companies' non-public training data without a contract or legal authorization. Preliminary public monitoring is possible using open data; in this case, all conclusions are limited to the available material.

The risk of independence	Minimum protection	Trace for verification
Payment by supplier	Fixed payment regardless of the result	Contract and tariff policy
Parallel counseling	Disclosure and recusal	Interest log
Pressure to publish	Independent Editorial Policy	Version and voting history
No closed access	Clear status desk review	Restrictions on title and conclusions

34. Application to legal analysis, investigation and sanctions

For ARGA, copyright audits arise when using training databases of judicial publications, reference books, investigative journalism, and commercial sanctions compliance databases. A distinction must be made between the legality of using a public government act as a source of facts and the right to an editorial database in which the act is provided with annotations, classifications, and paid search infrastructure. A lawyer's access license to the database does not necessarily cover mass copying for training a proprietary model. [7][21]

Evidence from specific client cases constitutes a separate category. Even when a lawyer or expert has legally obtained documents, uploading them to the global training system without a separate analysis of confidentiality, personal data, professional duties, and rights to the documents may be unacceptable. A distinction must be made between closed retrieval for a specific assignment and the use of the client as a source of ongoing training for a general service. For commercial sanctions databases, contractual prohibitions on mass retrieval, the right to process, and the publication of consolidated analytics are similarly reviewed.

ARGA proposes to launch the pilot only with anonymized/licensed training data and an evidentiary registry. This study makes no actual claims that third-party models are already being used by INTERPOL or other organizations for specific ARGA materials. This analytical project does not replace official permits and access requirements for the relevant information.

Material	The main risk	Acceptable pilot approach
Official court order	Publication terms and personal data	Analysis of public facts and a separate GDPR review

Material	The main risk	Acceptable pilot approach
Commercial legal framework	Database and contractual access	API Analysis License and Approval
Client documents	Secrecy, personal data and IP	Exclude from general education without separate assessment
Sanctioned commercial catalog	Aggregator rights and quality of recordings	Contract, audit of sources, error correction

35. Recommendations for consultation European Commission Technology and Copyright

The European Commission's targeted consultation, launched on 29 September 2026 and open until 3 November, provides a suitable forum for proposals on transparency of the copyright origin of training data. Responses should adhere to the framework of the questions in the form, rather than declaring submission in the report itself. ARGA's proposed position is a combination of verifiable transparency, effective rights reservation, accessible protections for small authors, and manageable conditions for access by independent auditors. [5]

The first proposal is to establish a compatible minimum set of provenance fields: source category, collection period, legal basis, confirmation of legal signal processing, model identifier, and modifications during additional training. The second is to develop machine-readable reservation mechanisms with free and non-discriminatory access. The third is to ensure secure access to sufficient evidence for legitimate control without requiring public disclosure of the entire licensed corpus. The fourth is to develop low-cost challenge procedures for individual copyright holders and independent testing protocols.

Proposals must be accompanied by an assessment of the costs to developers and rights holders. Unreasonably detailed public reporting can expose commercial contracts and infrastructure vulnerabilities; a report that is too general becomes useless for rights enforcement. A compromise: a public layer with sufficient categorization, a closed, verifiable registry, and legitimate independent verification procedures. Participation in the consultation is possible only after formal submission and confirmation; this study does not in itself constitute a submitted ARGA position.

Offer	Target	Proportionality condition
Provenance standard	Traceability of origin	Multi-level disclosure without a full dump
Machine-readable opt-out	Real enforceability of rights	Free interface for small authors
Evidence-based access	Possibility of independent verification	Secure environment, access log
Out-of-court consideration	Reducing defense costs	The right to appeal and access to court are retained

36. Implementation plan and quality indicators of independent control

The practical implementation program consists of three sequential blocks: a regulatory roadmap, a provenance registry, and an independent review of pilot datasets. In the first stage, the legal team creates a license classifier, a TDM, the legal status of models, and procedural rules. In the second stage, a technical specialist constructs a provenance graph for a small controlled dataset and tests opt-out processing. In the third stage, a separate team conducts a blind review of the sample and prepares a report with evidence, uncertainties, and the right to adjust the results.

Recommended indicators should not include the artificial goal of "zero complaints." It is more useful to measure the proportion of records with a traceable source, the proportion of confirmed legal grounds, the frequency of incorrect opt-out classifications, the response time to a credible complaint, the completeness of the manifest after an update, and the proportion of reports with independent review. For each indicator, the population, measurement method, interpretation limits, and escalation criteria should be specified.

A pilot project cannot be launched using confidential client data solely for portfolio expansion. Stop conditions must include the lack of a legal basis, proven circumvention of the ban, the inability to isolate the disputed dataset,

significant gaps in provenance, and an unmanageable conflict of interest. The final decision on publication should not be based on the report's page count: high-quality work may result in a motivated refusal to use the data.

Stage	The result being tested	Reason for stopping
Normative map	Matrix of responsibilities and statuses	Unspecified applicable law
Data Pilot	Origin graph and opt-out log	Significant unauthorized assets
Review	Reproducible sampling and output	Conflict of interest without protection
Publication	Reservations, sources, revision program	Unverified Certification Claims

37. Practical conclusions: what is confirmed and what remains controversial

WIPO's international materials further confirm that comparative assessment of AI and intellectual property must maintain a distinction between national exceptions, licenses and conditions for placing a product on the market. [14]

EU law contains clear, mandatory elements: the TDM exception under certain conditions, the general reservation right, the reproduction right, and the GPAI provider's obligations regarding copyright compliance and public disclosure. At the same time, many questions of judicial qualification are complex: how to apply a specific exception to different stages of training, when a legal signal is considered appropriate, how to prove the connection between a work and a model, and what factors determine civil liability for output coincidences. Final answers cannot be replaced by internal audit rules. [1][2][6]

The institutional conclusion is that transparency without evidentiary traceability is weaker than a rights protection mechanism, and a license without a chain of authority is weaker than genuine authorization from the copyright holder. However, one cannot demand impossible absolute certainty: a conscientious audit characterizes the limits of what has been verified and develops manageable risk mitigation measures. For ARGA, this provides a basis for professional research and participation in consultations without feigning regulatory authority.

This report contains legal methodology, practical forms, and hypothetical scenarios; it does not prove copyright infringement by a specific company or constitute an assessment of the compliance of its closed models. Any advancement to a factual conclusion requires a separate assignment, secure access, and a reproducible analysis using a documented protocol. Information on the status of contentious litigation should be verified as of the date of official publication.

Result	Status	Proposed ARGA Action
TDM exceptions and opt-outs	The current EU regulatory framework	Methodological explanations and comparative conclusions
GPAI Public Summary	Mandatory regulatory obligation where applicable	Checking the disclosure level and versions
General illegality/legality of education	Not established as a universal rule	Refusal of categorical conclusions
Provenance audit	Proposed independent methodology	Controlled body pilot
Consultation Technology and Copyright	The current public procedure runs until November 3, 2026.	Agreed documented position

38. Evidence of Market Damage: Causality, Substitution, Alternatives

When discussing damages from model training, it's important to distinguish between several possible events: the creation of an unauthorized copy, loss of license fees, reduced sales of the original work, a reduction in the value of a future license, and the emergence of a competing license. These events are not equivalent in terms of the nature of the evidence and possible remedies. For example, the absence of a license agreement prevents one from automatically multiplying the corpus size by the maximum commercial rate and calling the result established damages. The subject matter of the agreement, the actual market, the timing, the permissibility of licensing, and applicable national law are all relevant. [10][13]

When studying competition, it's useful to construct a counterfactual scenario: what would have happened without the controversial technology, but taking into account other factors influencing demand. For photographs, this

could be a change in platform distribution and demand for stock images; for legal annotations, the development of free official databases; for music lyrics, the growth or decline of licensing services. This assessment requires data on prices, transactions, and the quality of substitutes, not just a survey of stakeholders. A correlation between declining sales and the rise of AI cannot be presented as proven causation.

A three-layered assessment is proposed: a legal qualification of the claimed right; an economic model of the alleged harm; and an independent sensitivity of the result to alternative explanations. The value of the works themselves must be specifically separated from the value of commercial collections, interfaces, markup, and service infrastructure. For the sake of transparency, restrictions on access to commercial data and the source of such data must be disclosed. Even a sound economic model is no substitute for a judicial assessment of the relevant circumstances.

Harm hypothesis	Required data	Risk of error
Unreceived license	Comparable contracts and rights	There was no market for this license.
Loss of sales	Sales, demand, substitutes, time	Blending with industry trends
Substitution of the original service	Functional similarity and user behavior	Similarity of task does not mean substitution
Costs of protection	Real costs and causality	Attributing normal expenses to an incident

39. Royalties and Collective Distribution: Verifiable Arithmetic

Compensation models can be based on fixed payments, volume of works, intensity of use, economic contribution assessment, or collective distribution. Each carries the risk of manipulation. In particular, the number of tokens or copies of a file can be highly dependent on technical segmentation and duplication and does not reflect creative value; reviewing model outputs does not always allow for reliable determination of the causal contribution of a specific work. Therefore, the proposed formula must be verifiable and accompanied by disclaimers regarding the impossibility of precise attribution of each unit of generation.

A case study: a license pool of €120,000 leaves €108,000 after the agreed-upon administrative withholding of 10%. If 60% of the amount is distributed proportionally to the confirmed repertoire, and 40% according to objectively determined classes of use, the corresponding shares are €64,800 and €43,200. This is a mathematical illustration, not an existing tariff or a proposal for mandatory withholding. Actual criteria should be tested for discrimination against small rights holders, cost transparency, and the right to object.

An audit is important not only to verify the percentage multiplication but also to ensure the legitimacy of the works included in the numerator and denominator. Duplicate copies, unverified company credentials, and catalogs outside the territorial scope can systematically distort payments. It is advisable to establish public rules, a closed, auditable accrual statement, and an independent mechanism for correcting misattributions.

Step	Training calculation	Limitation
Source pool	120,000 euros	Notional amount
10% retention	12,000 euros	Only if agreed upon
Clean pool	108,000 euros	Check the register of copyright holders
Class A (60%)	64,800 euros	The weight of the repertoire requires justification
Class B (40%)	43,200 euros	The weight of use requires evidence

40. Accuracy of legal reservation: time and legal effect

Conflicts often arise not from a complete lack of signal, but from a temporal mismatch. The copyright holder issues a machine-generated injunction after the material has previously been copied and used in one version of the model; the developer later creates a new training set from the old copy. Legal assessment requires four dates: initial access, copy creation, reservation, and new use. The entire process cannot be viewed as a single, eternal action, as different operations can have independent legal qualifications. [1][21]

The audit system must store the site's terms and conditions at the time of each download and the legal decision regarding subsequent use. If a signal was hidden due to a technical error by the crawler, it must be determined whether it could reasonably have been detected and whether the mechanism complies with current technology. If an author mistakenly attributes someone else's work, a verification of authorship is required, rather than automatic

permanent blocking of all intellectual activity. In the case of multiple copies, a matrix of the specific instances and derivative sets affected is required.

ARGA's recommendations should include a procedure for the voluntary resolution of historical disputes without prejudging retroactivity across all legal systems. This could include a voluntary license, a limitation on future collection, technical measures to prevent release, or an agreement on the use of the disputed area. However, a legal position on past conduct should be based on law and evidence, not on the convenience of a company's policy.

Event	Fixed date	Legal issue
First legal access	T1	What was the basis?
Creating a copy	T2	Was copying allowed?
The emergence of opt-out	T3	Is the form and addressee correct?
Refresher training	T4	Is a new legal assessment required?

41. Trade secrets and closed audit: the limits of transparency

A full public list of works, source URLs, the value of each license agreement, and the structure of the dataset may affect protected trade secrets, security, and third-party rights. However, this argument does not equate to a right to waive all proof of origin. The AI Act already establishes a certain public minimum summary, while simultaneously providing secure access routes to other documentation for competent authorities. A research proposal could develop a similar multi-layered architecture without presenting it as a mandatory, detailed inventory of every file. [2][3]

The public layer must clearly identify the types of sources, their timeframe, changes, and the procedures for enforcing copyright restrictions. In the confidential layer, an independent verifier with legal authority can examine specific contracts and records. For the judicial layer, the procedural safeguards of the relevant country apply. The goal is not for anyone to download the entire corpus, but rather to be able to verify a material assertion through a controlled and reproducible procedure.

Implementation requires an access policy, request registration, prohibition of unauthorized removal of materials, restrictions on the pool of experts, and separation of roles. However, a closed environment alone does not ensure the independence of the conclusion: the reviewer must be able to select elements of the sample and verify the completeness of the population, otherwise the client may only show pre-selected "good" records.

Level of disclosure	Material	Fuse
Public	Categories and process	Do not disclose other people's contracts
Confidential audit	Asset registry and license selection	Roles, Logs, Independent Selection
Regulatory	Documentation upon legal request	Competence and protection of confidentiality
Judicial	Materials within the process	Procedural guarantees of the parties

42. Proof of Origin Format: Compatibility and Portability

Lineage history is only of practical value if it survives changes in technologies and vendors. A registry tied entirely to a platform's proprietary name will complicate migration between cloud environments and independent verification. It is recommended to record universal attributes: object ID, format, hash, source, time, legal basis, consent/objection status, parent-child dependencies, change operator, and the identifier of the model version used. At the same time, metadata standards themselves should evolve with the participation of authors, developers, and auditors.

The provenance of training data should not be confused with the provenance of the generated output. Metadata confirming which camera captured a photograph can help establish the origin of the original, but does not automatically prove that this particular photograph was present in the training corpus. Similarly, a digital signature of a public summary confirms the integrity of the document, but not the veracity of each statement. Strong inference requires linking verified technical records with legal approvals.

The system must correctly account for transformations that change the hash but leave the protected work recognizable. These cases require approximate matching algorithms, false-match checking, and preservation of

original evidence. If a company claims high accuracy in recognizing the identity of works, the auditor should request data on the validation set and independent error assessment.

Provenance statement	Proof	What it does not confirm
"File received on date T"	Signed log and source	Legality of use
"It's the same piece."	Hash + content matching	All rights reserved for other versions
"There is a license"	Agreement and powers	No other restrictions
"The summary has not changed."	Digital signature	Veracity of content

43. Output Works: Authorship, Derivatives, Licenses

This report primarily addresses input training materials, but economic risk often manifests itself in generation. It is necessary to distinguish between the issue of authorship of a completely automatically generated result, the issue of protecting human creative additions, and the issue of the possible reproduction of protected elements of the original works. The absence of copyright on an individual purely machine-generated result under a specific right does not mean that copying someone else's protected form with that result becomes permitted. [6][10]

Therefore, product policies must distinguish between the terms of ownership/use of outputs, notices of possible third-party rights, and matching tests. An image generator can create a result resembling a general artistic style, but this is not the same as reproducing a specific copyrighted image. Alternatively, the output may contain clearly recognizable elements of a specific work. The distinction is established by substantive expertise, not by marketing phrases like "style transfer" or "model creativity."

The supplier and integrator must determine in advance who will review result claims, who will retain evidence, and what access restrictions are permitted until the dispute is resolved. The report should separately note the impossibility of establishing origin based on a single output match: such information can be a strong signal, but requires verification of the training input chain and alternative causes for the match.

Legal issue	Object of verification	Mistaken identification
Protectability of the output	Human contribution and applicable law	"AI is always the author"
Productivity	Essential protected elements	"Similarity of topic = violation"
History of training	Provenance and magazines	"Similar output proves URL"
Responsibility	Supplier/User Actions	"Only the user always answers"

44. Application of the AI Act to GPAI modifications: supplier boundary

The status of the entity modifying the model matters for the assignment of responsibilities under Section 53 of the AI Act. Some fine-tuning operations may result in a downstream organization becoming a provider of the resulting general-purpose model, but not every user interface customization or prompt makes it so. The Commission provides coverage criteria and clarifies the procedure for documenting new training data when modifying a model. Therefore, an audit must begin with a technical and legal description of the modification.[2] [3][8]

For licensing documents, this means checking whether permission to perform additional training extends beyond the original provider. A downstream participant may have the right to use the API, but not the right to copy the training inputs of the primary model. When creating their own new corpus, they are responsible for verifying the rights to additional sources; relying on the old public summary of the base model is insufficient for new data.

It is recommended to maintain a split proof of concept: the base model set and version, the proprietary delta, the new legal status, the sources of the declared exceptions, and the current summary. When publishing, avoid double-counting common sources and clearly identify the reference base model. Individual system risk and documentation responsibilities are assessed when applicable, rather than automatically inferred from the use of a large API.

Type of change	Necessary verification	Document
Prompt policy	Doesn't it change the legal status?	Architectural description
Fine-tuning	Own corps and role	Delta manifest
New GPAI model	Article 53 and terms	Legal status memo
New public summary	Change coverage	Link to the basic and new versions

45. Verifiability of contracts without publication of personal and commercial data

To verify the licensing basis, the auditor often obtains a contract containing addresses, banking terms, author information, personal signatures, and business terms. It is necessary to preserve a sufficient portion of the legal information while simultaneously excluding unnecessary disclosure. A practical model allows for multiple representations of a single contract: the full original in a secure archive, an analytical extraction of the material terms, and a public record of the scope of the right without commercially sensitive amounts. [20]

Minimum legal extractables: authorized parties, license subject matter, territory, term, permitted operations, sublicenses, termination conditions, and limitations on model types. Excluding price data from a public document must not obscure a material prohibition on machine learning. The editor is responsible for determining which information affects the legal conclusion and which can be generalized without loss of verifiability.

The access verification is documented in writing: who saw the full contract, what version, for what purpose, and where the evidence is stored. The report indicates whether a selection of original signed documents was made or only company-provided summaries were used. This distinction is important for an honest assessment of trust and to prevent misleading claims of external certification.

Layer	Type of information	Control
Complete original	All conditions and signatures	Secure access
Audit extract	Essential rights and prohibitions	Check against the original
Public post	Category and range of law	No price/personal data leakage
Verification history	Who, when, and which version was seen?	Audit trail

46. Comparison of disclosed and actually verified indicators

A public summary might show that 30% of sources are from licensed catalogs, but such a metric is of little use without a defined unit of measurement: percentage of files, tokens, byte size, unique works, or license costs. Each category must have a vocabulary and calculation method. The auditor is obligated to verify that a few large collections do not mask a significant group of materials of uncertain origin. [3]

For example, if 1,000 unique works are represented by 10,000 files, and a license is confirmed for 900 works, the share of the licensed repertoire is 90% of the works, but does not necessarily equal the share of files. Extensive duplication of authorized materials can artificially inflate the "matching percentage" when calculating file-based calculations. Both figures should be shown, or the impossibility of one measurement should be explained, and the result should be tested for deduplication resistance.

The report should also disclose the availability of evidence. If 50% of sources have a documented legal basis, and for the remaining 50% no documentation is provided, it is inappropriate to describe the indicator as "50% legal, 50% illegal." It is more appropriate to record "50% evidence-based" and a separate category of "unverified," then identify legal and organizational measures to reduce uncertainty.

Indicator	Unit	Methodological limitation
Share of licensed repertoire	Unique works	Deduplication is needed
Share of licensed files	Files	Sensitive to copies
Provenance-complete share	Records with fields	Completeness of metadata does not equal legality
Audit-covered share	Parts of the population	Do not extend to inaccessible layers

47. Open publication and the right to object to an expert opinion

Publication of independent audit results can impact a company's reputation and the author's standing. Therefore, the procedure must ensure a clear distinction between facts, conclusions, and hypotheses, as well as a mechanism for correcting demonstrable errors. Before publishing materially negative statements, it is permissible to give the affected party a reasonable opportunity to respond to the facts being verified, provided editorial independence is maintained and security deadlines are met. This mechanism should not be transformed into a company's right to suppress inconvenient research.

The text must clearly indicate which data was obtained from public sources, which from the verified supplier, and which was verified by a third party. If a company disagrees with the legal classification, the report may contain a

brief statement of the company's position, but final responsibility for the methodological conclusion remains with the authors. Unsubstantiated accusations of "mass theft" are inappropriate, as are promises of "absolute compliance" if the researcher has not reviewed the entire corpus.

For Observatoire ARGA, it is proposed to separate actual corrections from changes in scientific assessment. New evidence should result in a dated supplement or a new version with a changelog, rather than a seamless edit of previously published material. At the end of each report, a contact information for corrections, the version of the methodology, the cutoff date for sources, and access restrictions are provided.

Event	Editorial measure	Maintaining independence
Factual refutation	Checking and adjusting	Publishing changes
Legal disagreement	Brief position of the party	The conclusion remains motivated
New source	Version update	Save the original archive
Sponsor pressure	Conflict Protocol	Recusal/independent decision

48. Double-check of Article 53 of the AI Act and substantive copyright

When preparing an audit engagement, two distinct lines of inquiry must be pursued. The first addresses the question of whether the GPAI provider meets its formal obligations under Article 53(1)(c)–(d): whether there is an applicable copyright compliance policy, how is the identification of rights reservation structured, whether a summary of the training content is published and linked to an identifiable version of the model. The second line concerns the lawfulness of specific actions on specific objects: whether lawful access can be demonstrated, who is the holder of the relevant rights, whether a TDM exception applies, whether there was a valid reservation, what is the territorial and temporal scope of the license, and whether separate reproduction occurred during generation. The first line does not replace the second. Article 53 does not create a general copyright license for the entire Internet, and a finding of lawful TDM does not relieve the provider of the obligation to be transparent where the AI Act applies. [1][2][3][8]

It's useful to construct an evidentiary matrix in which each regulatory requirement is matched by one or more sources, a responsible owner, and a sufficiency criterion. For the regulatory framework, this includes the approved version of the policy, the execution log, a public summary page, and a history of changes. For the material framework, this includes contracts, license terms as of the access date, crawler logs, opt-out signal snapshots, work identifiers, and rights chain information. If the verifier only has access to the public summary, the conclusion must be limited to its completeness and form: category-level transparency is not equivalent to confirming the legitimacy of each element of a multi-billion-dollar corpus.

A methodological error occurs when an organization introduces a single "95% compliance score" metric, mechanically averaging the presence of procedures and identified violations of exclusive rights. Confirmed illegal copying cannot be compensated for by a well-written policy. Therefore, the results are presented separately: procedural compliance, evidence of rights based on a sample, significant uncertainties, and decisions to continue or discontinue the use of disputed sources. The decision threshold is determined in advance, not chosen after unfavorable results are identified.

Circuit	Control statement	Proof	Ban on withdrawal
AI Act, policy	The opt-out mechanism is in effect and not described.	Versioned policy, audit logs, incidents	"Therefore, all works were used legally."
AI Act, summary	The summary covers the specified model/version	Public link, date, source categories	"The report details each individual object."
Copyright	There is a basis for specific reproduction	License or justification for exception and facts of access	This license covers all versions of the case.
Model output	No unauthorized significant reproduction was identified in the tests	Query protocol, comparison criteria, sampling restrictions	"Any future generation is safe"

49. Testing multimodal learning: image, sound, video

The same criteria for analyzing book texts cannot be applied to images, phonograms, and video materials without additional evidence. A photographic archive contains the rights of the author of the photograph and possible rights to the image; a musical recording links a work, a performance, and a phonogram; a video combines images,

sound, editing, and often fragments of third-party works. A single permission from the server owner cannot automatically cover all these layers. Therefore, an audit of a multimodal corpus begins with a map of the types of rights and parties that potentially must grant permission. A publicly accessible file on the Internet does not become public domain due to the loss of EXIF metadata upon republication. [6][7][20]

The technical protocol contains not only a cryptographic hash of the file but also carefully tuned perceptual features. Exact hashes are reliable for bit-by-bit matching, but they don't link the original JPEG to the modified crop, while perceptual hashes can produce false matches. A similar problem arises with audio fingerprints, accelerated playback, or noise reduction. A fingerprint match does not establish the copyright holder; it merely points to an object for legal verification. When in doubt, manual verification is used with a recorded confidence level, rather than automatic conviction.

For output reproduction, specific model modes should be tested: text generation, image-to-image, audio continuation, voice imitation, and video with reference frames. The conditions for presenting the original model are documented especially carefully: if the researcher themselves provided a protected image or music fragment as reference, a similar output has a different evidentiary value than reproduction without providing the source material. Published illustrations should be small enough and legally usable, or replaced with a description of the result, to prevent the analytical report itself from becoming a channel for repeat infringement.

Object	What does a technical expert record?	What does the lawyer separately establish?
Photo	Transformations, crop similarities, source of copy	Attribution, license chain, usage limits
Sound recording	Fingerprint, tempo, transformations, presence of reference	Rights to a work, performance and phonogram
Video	Segments, duration, audio synchronization	A set of rights to composite objects
Synthetic voice	Technical sign of imitation	Related, personal and contractual rights under applicable law

50. Architecture for secure storage of evidence and litigation hold

Information collected during an audit may subsequently become evidence in court or subject to regulatory access. This does not mean it can be stored indefinitely. The retention policy sets a retention period based on the type of object, the basis for retention, the anticipated dispute, and the personal data regime. In a typical audit cycle, temporary copies are deleted or minimized after the purpose is achieved; in the event of actual or reasonably foreseeable litigation, the admissibility and scope of temporary retention are assessed separately with legal counsel. It is important to distinguish between the retention of minimal evidence and the continued use of the disputed work for educational purposes.

For critical elements, a package is generated: file source, protected identifier, checksum, header and contractual metadata extraction, proof of the web page state at the time of access, request logs, time and environment of receipt, transmission chain, and reasons for each copy. The research environment's clock is synchronized and time-stamped. Corrections to the protocol text are made by a separate entry, not by overwriting the original log. If the log was generated by a single participant without external verification, the degree of independence of such proof should be clearly stated.

Because source works may be commercially valuable or contain personal information, a public report requires a separate layer of anonymized evidence: aggregated numerical tables, secure links to entries in a closed registry, repeatable test parameters, and a list of untested hypotheses. A secure depository with an access log and destruction dates allows for subsequent verification without uncontrolled distribution of the full corpus. An access protocol separates the client's employees from independent reviewers and records the disclosure of potential conflicts of interest.

51. Contractual warranties, audit rights and limits of assurances

When purchasing a training set or a commercial license, it's important to check not the aesthetics of the warranty formula, but its legal applicability. A vendor's claim of "all necessary rights" may be limited by a contractually agreed-upon list of objects, territory, maximum volume of use, content acquisition date, or the right to distribute derivative datasets. Essential rights include machine learning, the right to make copies and transformations,

subcontracted use, further training, weight preservation, re-versions of the model, use of fragments in output results, and the transfer of evidence to the auditor. None of these permissions should be automatically implied in the absence of a corresponding provision or a clear contract.

The agreement, designed for subsequent audit, includes a manifest of licensed resources, a procedure for updating the list, notions of origin, deadlines for submitting notifications of claims, document formats, the right to a limited audit, and a procedure for challenging findings. Compensation for damages and the limit of liability are calculated separately from the obligation to provide access to evidence. The prohibition on providing unsubstantiated source files to third parties does not preclude the use of a secure data room, where the auditor is permitted to review a control sample without copying the entire dataset.

It's impossible to promise that a single intermediary's license will protect against the claims of all unknown authors. For complex registries, the developer requires documentation of the intermediary's authority and rights distribution methods. If the source was actually obtained through an aggregator that only has public display rights and not permission to reproduce for educational purposes, the auditor will document the specific gap and possible remedies: an additional license, exclusion of the disputed catalog from future cycles, clarification of the rights chain, or a legal assessment of the applicable exception. Payment to the provider alone does not resolve the copyright issue.

Contractual issue	Minimum evidentiary form	Consequence of absence
Who granted the rights?	Powers of the parties and map of copyright holders	Unable to confirm chain of title
For which versions and territories?	List, license, temporal scope	Output only for the committed version
Is independent testing possible?	Limited access, NDA and journal	Significant limitation of the conclusion
How to respond to complaints?	Notification and Correction Procedure	Increased operational risk

52. Legal dimension of open repositories and software code

Publishing software code in an open repository does not equate to a waiver of copyright. An open license defines the scope of permissions and associated obligations, such as maintaining notices, disclosing modifications, or distributing derivative works under specified terms. Training datasets present a distinct problem: nested dependencies, unlicensed files, copies of third-party libraries, and code contributed by unauthorized contributors. A policy like "the repository is marked as open source, therefore all code is free" doesn't hold up even to the basic assumptions.

The audit is conducted layer by layer: the project license, submodule licenses, the content of vendored dependencies, the temporary version, access conditions, and the legal classification of actions. It should not be automatically assumed that all training on code creates a derivative work of software or, conversely, that any reproduction in the model is impossible. These are different issues that require technical verification and applicable law. Commercial use of open source code may be permissible if the license terms are followed, but the terms themselves cannot be ignored simply because the code appears in a dataset.

To test the code generator, a control sample of rare functions, signatures, comments, and licensed notification texts is used. Similarity between short, standard constructs has a different evidentiary value than the reproduction of a long, unique fragment; the comparative methodology takes into account functional necessity, the standard nature of the solution, and available prior code. The final assessment should separately evaluate the risk of releasing proprietary code, the risk of non-compliance with license terms, and the risk of vulnerabilities in the generated code. The latter does not automatically constitute copyright infringement.

53. Photo provenance, metadata, and control of false attributions

For photo archives and professional illustrations, a common problem is not only lack of licensing but also misattribution. An image may appear on an aggregator without an author's signature, with fake metadata, or under a changed filename. A direct link to an aggregator does not establish a chain of rights. The researcher identifies the earliest available verifiable source and compares it with licenses, transfer deeds, agency contracts, and archival publications; if there is conflicting evidence, the conclusion remains preliminary. [6][20]

Technical comparison combines precise hashes, perceptual similarity, frame matching, captions, and publication history, but does not derive a "98% probability of authorship" from a single metric of visual similarity. For

retouched, cropped, and recompressed images, it is necessary to demonstrate which features remain. The terms of the platform on which the photograph first appeared are verified against the historical date. Legal display of a photograph by the copyright holder does not imply permission for the automated download of the entire library.

A separate registry of images that cannot be included in their entirety in the public version of the report is maintained. If necessary, the expert is provided with a protected identifier and a brief description of the match, along with a minimum illustration within legally permissible limits. For portfolios and concert footage, author attribution is particularly important: an erroneous caption can have not only financial but also reputational consequences. Therefore, corrective action encompasses rights, credit approval, and correction of published metadata, rather than simply deleting a single copy of the file.

54. Music and Related Rights: Different Layers of Permissions

Training a model on a song may involve simultaneously the protected musical work, lyrics, performance, and a specific phonogram. Rights may belong to different entities; managing a portion of a repertoire through a collective organization does not establish authority to permit machine learning of all its layers. Therefore, a legal request to a music catalog must separately list the objects, territory, term, permitted uses, and the possibility of modification and transfer of derivative information. Checking a streaming license alone does not establish permission to make training copies or use the recording to configure a generator. [6][19]

A separate technical scenario concerns the detection of long, verbatim fragments of song lyrics, and another concerns the similarity of a new audio recording to a specific phonogram. The legal criteria for these differ. A short sequence of common words does not have the same evidentiary force as a characteristic protected text fragment, and stylistic similarity of music does not equate to the identity of the protected recording. The expert describes the objective features and the method of comparison; the legal qualification takes into account originality, content, and circumstances. The Munich court's November 2025 decision in the GEMA dispute is relevant for the discussion of memorization and reproduction, but does not constitute a universal final rule for all European courts. [11]

The report should distinguish between copyright risk, the artist's right to vote and personal information, requirements for labeling synthetic material, and possible contractual prohibitions on imitation. Attempting to lump all these issues under the single term "copyright violation" makes the report unsuitable for court review. It is recommended to formulate separate assessments by rights layer, with a clear chain of evidence and specific targets for potential corrective measures.

55. Publishing archive, digital libraries and access licenses

In relation to publishing archives, it is essential to separate the rights to read content, reproduce it in the research process, extract a significant portion of the database, and use the material for a commercial generative model. A subscription or open search engine listing does not equate to a transfer of exclusive rights. In publishing catalogues, the rights of individual authors and photographers, as well as the rights of the database producer and the special rights of the press publisher, are possible. A legal opinion requires a map of objects linked to specific articles, versions, and retrieval dates. [6][7]

The audit records the type of access: public page, subscription, institutional license, legitimate purchase, aggregator, or mirror. A research organization's academic status does not transform any commercial educational task into research TDM under Article 3 of the DSM. When assessing Article 4, it is necessary to check the legal requirements and any reservation of rights. The assumption that every robots.txt file is equally opt-out is too crude: the signal content, terms of use, and technological circumstances are what matter. The decision is made based on specific facts, not a binary tag.

It is proposed to use a protected pilot sample of a limited number of publishers and access types, recording in advance the percentage of missing licenses, attribution conflicts, source blocking, and claim review times. When dealing with personal data, copyright and GDPR compliance are checked in parallel. The legality of personal data processing cannot be inferred solely from the presence of a copyright license, nor can the opposite conclusion be drawn without analyzing the basis for the processing.

56. Representativeness of tests: why negative controls are needed

The memorization test protocol can be easily skewed by selecting only the most recognizable passages or repeatedly trying prompts until a match is found. Such an experiment demonstrates the feasibility of reproducibility under search conditions, but not the prevalence of the phenomenon in everyday use. Different questions require different populations: all output responses, all works in a specific catalog, all queries with a given context, or all versions of a single model. The denominator is declared before the test. The results of intentional retrieval, natural use, and repeated queries are considered separately.

Negative control includes texts and phrases missing from the analyzed corpus for reasons known to the auditor, as well as standard, unoriginal phrases. It helps assess false method matches. Positive control may include pre-approved and legally used test pieces; however, its presence does not prove the model's performance on other subjects. Several researchers separately mark the significance of matches; disagreements are reviewed on a journal-by-journal basis rather than being hidden by an average result.

The result "0 matches found out of 300 attempts" does not imply zero probability of leakage. Under independence assumptions, one can report an approximate upper 95% bound of 3/300, or about 1%, but if the queries are correlated, such an estimate has no such interpretation. It is equally unacceptable to interpret "12 matches out of 300" as the proportion of all illegally used works in the training. The technical test answers a narrow question about observed search results; a legal audit of the corpus requires a different sample.

Observation	What can be concluded?	What cannot be concluded
Match after reference input	The model is capable of transforming the provided object	That she kept this object in training
Match without reference	A reproducible risk was detected under test conditions	That all models of this class violate the rights
0/300 matches	There are no matches in this protocol.	That the risk is zero
Similar style	Stylistic similarities have been identified	That the protected form must be reproduced

57. Decision Points: Stop, Correct, Publish

A good independent audit concludes not with a risk index, but with a limited set of decisions made with responsible parties. The situation may require additional documentation requests, restrictions on the use of the disputed dataset, obtaining a license, correcting the public summary, adjusting search filters, or transferring materials to legal counsel. For each decision, the legal basis, unverified assumptions, deadlines, and closure criteria are specified. Priority is determined not only by the likelihood of a dispute but also by the scale of potential damage, the nature of the work, the possibility of repeated copying, and the presence of urgent injunctions.

Escalation should not be based on a presumption of guilt. If the source has not provided a license to the auditor, this limits the evidence but does not constitute an admission of wrongdoing. If a contract is discovered but only covers the demonstration of the material, a documentable gap in the training justification arises. If a rights reservation is obtained after a certain date, the consequences for new collections and existing copies must be assessed separately; a late signal cannot be automatically transferred to all previous actions without legal analysis. [1][21]

A division of roles is being introduced for the independent Observatoire: the researcher establishes the facts, the legal expert qualifies the risks, the editor evaluates the safe public form, and the responsible person resolves disputes over interests and resources. The person conducting a commercial dispute with the audited company should not unilaterally approve the public conclusion regarding it. The public announcement includes the version and date of the methodology, a clear description of the limitations, the right of the audited party to submit documents, and a mechanism for correcting a material error.

58. Substantive review of EU consultation proposals

In the 2026 Technology and Copyright consultation, it is appropriate to distinguish between measures that can be implemented voluntarily and proposals that require changes to the regulatory framework. The former include a standardized format for consent logs, a unified, limited data room for independent auditors, machine-readable opt-

out registration methods, and a description of the origin of sources to the extent that allows for the protection of trade secrets. The latter may concern procedural access to evidence, compensation mechanisms, and pan-European rules that cannot be imposed on companies by a single methodological report. The consultation should not be presented as a final regulatory act. [5]

Any transparency proposal must be assessed for proportionality: disclosing a full list of URLs can sometimes facilitate the illegal downloading of archives, revealing personal data, or trade secrets, while publishing only the general term "Internet" deprives copyright holders of a verification tool. A viable compromise is public categories and aggregated shares, plus reasonable, closed access for competent authorities and independent experts, with strict restrictions on distribution. The verification method must be replicable for smaller copyright holders, otherwise the innovation becomes a paid pass for large catalogs.

Suggested metrics for future evaluation include: the proportion of sources with confirmed legal basis, the median opt-out time, the number of corrected attribution errors, the proportion of claims with reviewed evidence, the errors of the automatic rights classifier, and the proportion of models with up-to-date summaries. When publishing, the denominators and observation period should be recorded; otherwise, comparisons between participants will be misleading. Before submitting a position to ARGA, it is important to agree on the mandate, institutional status, the degree of representativeness of the findings, and the inadmissibility of presenting the proposed methodology as an already conducted empirical study.

59. A final model for assessing evidence without false precision

A four-level evidentiary certainty scale is useful for the practical organization of the Observatoire's work. Level one includes only supplier assertions and public marketing materials. Level two includes verifiable documents and sample logs, but the audit did not have independent access to the corpus. Level three provides secure control samples, reproducible tests, and independent document comparisons. Level four allows for extensive external verification of relevant lifecycle stages with sufficient completeness, time-bound coverage, and independent reviewers. The level numbers describe the completeness of the evidence but do not replace legal qualifications.

Publication should not turn the absence of error in the available sample into a "certificate of legality" for the entire model. Any issue found should be tied to a specific object, legal regime, stage of use, period, and level of proof. The factual question of "was the file used" is not the same as the legal question of "was it prohibited?" The absence of a commercial license does not preclude the application of a statutory exception; the presence of a license in one jurisdiction does not guarantee its universal applicability abroad. [1][2][6]

Recommendations for payments and corrective measures may be included only as proposals with a basis and competence, and not as binding decisions of a private expert organization. The final structure of the report must allow a third party to replicate the review process: the list of materials, the sampling decision, the results, the alternative hypotheses, the limitations, the dissenting opinions, and the final actions. This evidentiary discipline builds trust in the Observatoire better than the artificial "98.7% legal AI" rating without any clear meaning.

Level	The nature of the evidence	Formulating a public conclusion
1	Self-declaration and open pages	"The supplier stated that there was no independent verification."
2	Contracts and partial journals	"Some of the procedures are documented; there are limitations."
3	Control sampling and repeatable tests	"The specified sample, versions, and criteria have been verified."
4	Enhanced independent access and peer review	"Specific elements in the established volume were tested, not the entire future release."

60. Specific exceptions: quoting, temporary copying, and TDM

Discussions about model training often conflate several legally distinct situations: technical copying during data transmission, data mining for automated analysis, reproduction of a limited quotation in the text of a scientific study, and distribution of an extended fragment of a protected work to a user. Each case has its own specific conditions. In particular, the TDM exceptions in Articles 3 and 4 of Directive 2019/790 apply to reproduction or

mining for the purposes of text and data mining and do not constitute a universal license for all future forms of dissemination of the original work. [1][6]

When assessing an exception, it is necessary to examine, in a sequential manner, the lawful access, the type of activity, the user category and purposes, the existence of a reservation under Article 4, the copy retention requirements, and the connection of the disputed use with applicable national law. The general right of citation, where applicable, has its own conditions and does not replace this review. The technical fact of posting the source in a search index does not prove that the creation of a full training copy fell within the relevant exception. Conversely, the existence of another legal license may make the analysis of the TDM terms not the sole basis.

In an independent opinion, recommendations should be operational: for each disputed action, select one of the categories "permitted by contract," "potentially covered by an exception if conditions are proven," "reason not supported," or "contradiction with a specific requirement established." These conditions should not be translated into mathematical scoring. If the legality of the result depends on a disputed interpretation, it is recommended to present the arguments of both parties and the limits of judicial practice, rather than imitate a definitive answer.

61. Collective Rights Management: Repertoire Accessibility and Mandate

Collective licensing potentially reduces transaction costs when working with large creative catalogs. However, the mere existence of a collective management organization does not guarantee its right to authorize training on all works in circulation, in all countries, and for all modes of use. Practical testing begins with its legal status, the scope of rights holders' powers, mutual representation agreements, subject matter repertoire, and revenue sharing rules. Proposing a new type of license must be distinct from asserting that such a license already exists and must be recognized by third parties.

For the generative model, the boundaries of indirect control are particularly important: composite works, musical texts and sound recordings, works with multiple authors, publisher catalogs, cross-border rights, and revocable mandates. If a license applies to only one layer of rights, this is specified separately. It is also necessary to examine the rightsholder's ability to claim usage reservations, the conditions for exiting the catalog, and the availability of an effective mechanism for contesting distributions. A research royalty model should explain why a particular calculation method is chosen and acknowledge cases where there is no observable causal relationship with the model's revenue.

The EU consultation proposals could describe a technically neutral registry of mandates and auditable payments, but they cannot assume the automatic creation of a pan-European mandatory scheme without a legislative basis. Small authors and performers must have comparable access to audit records of their repertoire; otherwise, the overall cash flow will be visible, while the actual distribution will be opaque. The cost of auditing must be separated from the rights to the works, and a right to correct an incorrect entry must be provided.

62. Press publishers' rights and editorial databases

The rights of a press publisher under EU law and the legal protection of a database constitute separate issues, alongside the rights of the author of a specific article. Even when a fact presented in a news story is not protected by copyright, the selection, editorial presentation, photography, and the significant size of the database may be protected. The developer should not classify an archive solely by the criterion "texts are searchable"; it is necessary to distinguish between sources acquired under a commercial agreement, subscription libraries, open official documents, and materials licensed under a Creative Commons license, including their possible limitations.

The practical case concerns a model trained on aggregated journalistic archives. For each collection, it is necessary to establish the supplier, the relationship with the publisher, the source of rights to photographs and author texts, and whether a specific extract falls under the established exemption. The request to the archive need not require the disclosure of all journalists and editors if a verified mandate for specific categories is sufficient. However, general assurances from the aggregator without the ability to verify at least a selection are insufficient for a positive independent conclusion.

The market for news archive licenses cannot be assessed solely by the subscription price for readers: the educational product may create a different demand and a different market. Contracts, usage trends, and alternative explanations are necessary to prove damages. When publishing research, citing lengthy copyrighted materials

should be avoided: titles, dates, identifiers, and aggregate estimates usually allow the method to be explained without reproducing the entire corpus. [6][7]

63. Institutional and scientific libraries: scientific status is not a general permission

The research TDM exception under Article 3 of the DSM is closely tied to the user's qualifications and purpose. The mere involvement of a university, library, or research organization does not automatically qualify any action by a contractor, commercial partner, or individual developer. The researcher must document the institutional role, the nature of the project, legal access, storage conditions, and the possibility of protected reproduction. If a research center transfers a created dataset to a commercial partner, this is a separate, legally significant transfer, not a technical detail.

A table of sources is prepared for internal verification: purchased e-books, scientific articles, open-access archives, materials with permissive licenses, and library scans. Each type may have its own restrictions, and the right to read a copy is not always the same as the right to transfer files to a third-party model. An incomplete agreement should not be supplemented with a fictitious permission for mass analysis. At the same time, the protection of scientific research requires that every legal technical copying not be converted into a mandatory commercial license without checking the applicable exception.

The methodological recommendation is to transparently separate academic review from custom commercial development. When a mixed infrastructure is used, the journal records who initiated the extraction, at whose expense, for what purpose, who received a copy, whether the partner was able to reuse the data, and who is responsible for subsequent deletion. A negative conclusion should be linked to a missing legal requirement, not to a subjective assessment of the commercial nature of the AI.

64. Reliability of metadata, provenance markers and machine signatures

Content origin metadata is useful, but by its very nature, it's a statement that must be verified. A signature attached to a file may certify that the file has been modified since it was signed with a specific key; it doesn't, by itself, establish that the signatory held all exclusive rights or had permission to use someone else's work. Similarly, standard creator metadata can be incomplete, copied, or outdated. Therefore, auditing combines technical evidence of integrity with contractual and organizational documents.

When constructing a provenance graph, the system must record both positive and negative information: who signed what object, at what stage of the chain, in what version, which fields are independently verified, and what event triggered the change. If an image loses some metadata after conversion, the missing fields do not prove the absence of the author. Alternative methods of verification must be preserved—contracts, primary publications, archival receipts, registries, and transfer certificates.

Mandatory publication of the full graph is not always proportionate: some records may contain personal data, confidential commercial terms, and internal identifiers. A standardized description of evidence classes and an accessible mechanism for independent verification with differentiated access rights are sufficient for public verification. The method for describing provenance should not become a technological monopoly for a single solution provider.

Label	Evidentiary value	Limitation
Cryptographic signature	Integrity with respect to the signed state	Does not confirm all exclusive rights
EXIF/author's signature	Possible source of attribution	May be changed or absent
Contract/transfer deed	The scope of agreed rights	The powers and limits need to be checked.
Download log	Time and path of file appearance	Does not guarantee the legality of copying

65. Deduplication and Aggregation: The Probative Cost of Losing the Original Trace

Preparing training sets often involves deduplication, filtering, and removal of original URLs. These processes are beneficial for model quality, but can undermine subsequent legal verification. If the system stores only one "representative" of a group of similar objects, the auditor must know which originals, licenses, and opt-out signals are included in the group. Otherwise, a legitimate copy could obscure a second one obtained from an unauthorized

source, or vice versa: a questionable source could mistakenly color the entire group of legally licensed versions red.

The technical scheme includes an index of source identifiers, a date and grouping method, a rule for selecting a representative file, and a mechanism for restoring provenance without preserving all the original data in the open zone. Each merge, split, and delete must have a version event. During recompression or normalization of texts, transformations that may affect identity and authorship verification are recorded. Matching by rough fingerprints is accompanied by a log of false positives and false negatives.

An analytical report should not state "the specified work is not in the corpus" if, in fact, only a single hash was searched, and all transformed variants were not verified. Similarly, it is unacceptable to say "the work was illegally used" when only a similar file without the established copyright was found. For each result, the indexing completeness and verified transformations should be indicated so that the legal expert understands the evidentiary limits of the technical procedure.

66. International versions of the work and the right of translation

Translations, adaptations, abridged publications, and localized editions may have their own protected objects and rights chains. A license for an original novel does not always cover a third-party translation, and a license for one language and market does not necessarily extend to the entire multilingual training corpus. Automated analysis requires distinguishing between the authorship of the original work and the translator's contribution, the existence of copyright exceptions, and specific contractual restrictions.

Example: a training set contains the original English article, a French translation by the publisher, and a machine translation by a contractor. Each element is identified by its source, date, method of creation, and the legal role of the contributor. It cannot be automatically assumed that a machine translation is free of copyright to the original; equally, it is incorrect to consider any automatic reworking an independent protected work without analysis. The summary column separately shows the origin of the content, creative contribution, licensing permissions, and dependency on the original version.

When auditing the model's results, it is important to determine in advance what is being tested: a verbatim reproduction of the original, a translation of a protected fragment, a retelling of a fact, or a separate creative adaptation. The similarity method in one language is not mechanically applied to another; materiality criteria are established by specialists familiar with the relevant legal and linguistic context.

67. Handling corrections and deletions without false technical promises

Responses to copyright holders often contain contradictory statements: "the original file has been removed from the repository" and "the model no longer contains information from the work." The former can be confirmed by the operational log, but the latter requires separate analysis and often cannot be universally proven. The weights of a trained model are not a file catalog, so simply removing the original file from the future training set is not the same as re-applying the training. Output filters, modification of the future training set, technical unlearning procedures, or retraining are all possible, but each measure should be described based on the actual result.

The auditor records the date of the complaint, the relevant model version, the list of terminated downloads, the applied security filters, the tests performed, the remaining risks, and the retest plan. If information has been excluded from future versions, this should be stated without implying the complete impossibility of any generation. The model may reproduce a similar form due to other sources or the general nature of the expression; this result must be distinguished from proven preservation of a specific source file.

The document for the copyright holder must list feasible measures, as well as available avenues for further legal challenge. The response to a copyright claim and the consideration of a data subject's request under the GDPR have different grounds and timeframes; combining them into a single, indistinguishable "deletion procedure" creates the risk of an inappropriate response. [20]

68. The Economics of Compliance for Small Providers and Authors

Not all participants can afford to maintain an in-house legal department, audit team, and their own license repository. This has implications for the assessment of future transparency mechanisms: an overly expensive regime may formally require the same actions from both the international model and the small developer, but in reality exclude the latter from the market. At the same time, low costs do not create a general right to use others' works without complying with the law. Therefore, the research model must consider the proportionality of tools without reducing the necessary level of protection for copyright holders.

The proposed minimum package for a small provider includes: a versioned source registry, a standard machine-readable reservation parser, a license change log, a public description of training categories, and a copyright holder appeal procedure. For independent evaluation, a stratified, limited audit with a transparent uncertainty range is acceptable; costly verification of the entire corpus should not be mandated by law unless required. When evidence is in short supply, publication of findings should be more cautious, not more bold.

For small authors, an accessible, free opt-out channel, a clear response, and the ability to check the status of their specific work without paying a commercial intermediary are important. A standard complaints format may be among the measures discussed in the EU consultation, but its use should not block other legally permissible forms of communication. Quality indicators should include both the cost of response and the number of legitimate complaints satisfied.

69. Challenging an expert opinion and procedural fairness

An independent investigation alleging possible violations by a particular company may impact its reputation and legal standing. Therefore, before public disclosure, there must be a right to correct factual errors, a reasonable process for providing relevant documents, and an independent mechanism for resolving conflicts of interest. The right to submit objections does not imply the right of the audited party to veto the research conclusion. However, failure to respond cannot be converted into an admission of guilt; the wording must specify which requests were made and which documents were actually received.

For editorial review, a list of assertions with their respective levels of evidence is compiled. Facts, sample conclusions, expert opinions, and regulatory proposals are separately noted. Each material fact is accompanied by a reference to the evidence, and each legal conclusion is accompanied by the applicable law and the date of the review. If one expert considers the evidence of a violation sufficient, while another disagrees, the reasons for the disagreement are reflected in the review and are not concealed under the guise of "collective unanimity."

A public correction must leave a history of the correction: which version was published, what was changed, and why. Silently replacing a PDF without a changelog undermines trust in an independent organization, especially in discussions about rights and tech corporations. These procedures also apply to cases where the research is commissioned by a potential plaintiff; the funding and role of the commissioner must be disclosed to the extent permitted.

70. Evidence scale for legal advice and court material

A working report for an academic audience and procedural evidence are not the same. In a scientific text, it is permissible to formulate a hypothesis, note the sample limits, and propose a test; a legal case requires adherence to the admissibility and presentation standards of a particular jurisdiction. An Observatoire expert should not promise that their report will automatically be accepted by a court or regulator, and cannot substitute for an attorney authorized to represent the party. A notarized screenshot also does not resolve the issue of the substantive legality of the content.

For potential legal use, a provenance package is required: who received it, when, on what system, what changes were made, how integrity was confirmed, and who had access. Experimental methods are documented prior to testing, including query selection criteria and types of control observations. It must be possible to reproduce the relevant calculations without infringing the exclusive rights to the test objects themselves.

The outcome may include four options: sufficiency for the research assertion, the need for further verification, the impossibility of a reliable conclusion, and a direct refutation of the initial hypothesis. None of these constitutes a

procedural decision in and of itself. In a cross-border case, the qualification of the evidence and the method of obtaining it require separate local legal assessment.

71. Consolidated set of minimum measures for the ARGA Observatory

The practical implementation of the report requires three independent functions: legal research, technical audit, and editorial publication. The legal function verifies regulations, licenses, the applicability of exemptions, and competence; the technical function verifies the origin of sources, model versions, reproducibility, and journals; and the editorial function verifies the proportionality of disclosure, the absence of unsubstantiated allegations, reputational risks, and the possibility of correction. Relying on a single specialist to perform all these functions without peer review is especially risky if the organization simultaneously provides paid consultations to stakeholders.

For a pilot, it is recommended to limit the subject matter to a few types of data and sources. For example, one category of publishing materials, one group of photo archives, and one collection of open source code allow for testing different rights chains within a manageable scope. Before the launch, the permitted data access format, NDA conditions, statistical selection criteria, evidence ledger format, and public disclosure limits are specified. Unexplored categories are explicitly identified in the report; the pilot should not represent certification of the entire model.

The final document should provide management with a substantive decision: which procedures can be launched immediately based on existing law and voluntary agreements; which measures require a separate agreement or resource; which proposals should be submitted for public consultation; and which ideas remain research hypotheses. Accountability in the field of copyright does not consist of a maximum number of tables, but rather in the ability to verify every significant conclusion and correct errors in a timely manner.

Practical applications and research protocols

Appendix 1. Passport of the training data corpus

The hull passport is a high-level document that allows for establishing responsibility for the set and identifying which model versions received materials. It is completed before first use and updated whenever there is a significant change in the composition or authorized purpose. A single passport for all company hulls is inappropriate if it mixes licensed data, TDM sources, and proprietary client information. The minimum unit of record is a managed set with its own provenance and purpose.

The card must contain the owner, legal entity, contact for copyright holder inquiries, creation date, collection date range, model and version, intended use, upstream supplier registry, risk classification, and approval status. A separate field records unknown sources and reasons for incomplete information. The signatory confirms the completeness of the evidence provided, not absolute freedom of all rights worldwide. When expanding the use of the kit for a new model, a re-legal assessment is conducted.

Passport field	Recording format	Verification rule
Dataset ID and version	Unique code, date, responsible	Communication with manifest and backup
Purpose	Pre-training / fine-tuning / evaluation / RAG	Prohibition of other purposes without review
Sources	License / DSM3 / DSM4 / other basis	Mandatory link to the registry
Risks	By category of law and supplier	Leave a separate "unknown" flag
Solution	Admission / conditional / quarantine	Signature and date, version of the regulations

Appendix 2. Certificate of acceptance of a new source

Source acceptance begins with determining the actual owner of the acquisition process. It's impossible to accept a million records for a single "open source" claim without verifying the legal basis, machine-readable restrictions,

and reuse rights. The acceptance report must record the exact domains, final URLs after redirects, crawl period, content category, authentication method, license terms, and technical means of collection.

Acceptance is conducted in two independent stages. A technical specialist confirms that the crawl log corresponds to real objects and processes signals correctly. A legal specialist decides whether a specific license or exception can be applied. In the event of a dispute, access is restricted until a resolution is reached. Data for which it is impossible to determine its origin or for which prohibited sources are clearly indicated should not be included in the training simply because it has already been downloaded.

Examination	Conditions for passing	Document
Identity	The source and supplier have been identified	Registration card
Access	Access permission confirmed	Conditions on the date of collection
Rights	The basis is related to the type of operation	IP Memorandum
Robots/opt-out	The rules have been worked out as of the effective date	Machine log and its hash
Confidentiality	Prohibited categories are excluded	Dual IP/GDPR assessment

Appendix 3. Log of rights reservation and opt-out processing

The log is needed to answer four questions: who expressed their intent, in what form, when the signal became available, and which assets were excluded or further verified. For each event, an immutable record is created containing a snapshot of the original signal, the interpreter version, the authority verification status, and the result. The system must work with both automated website files and individual, motivated requests, without confounding their legal consequences.

If a signal is detected after the initial copy, the journal should not automatically overwrite the previous history. The initial access date and the status change date are recorded separately for future operations. If conditions change, a legal review of the appropriate follow-up actions is conducted. Collisions between mirror copies, author pseudonyms, syndication, and the rights of multiple co-authors are separately recorded. Notification of the result allows the submitter to verify that their request has not been lost in the automated system.

Event	Required data	Solution
Machine-readable signal	URL, timestamp, snapshot, parser version	Accepted / requires review
Author's notice	Object, authority, contact	Checking rights and scope
Conflict of rights	Multiple applicants	Freezing of the disputed category
Downstream route	Recipient Sets/Models	Exclusion from future builds
Closing	Motivation, date, responsible person	Send a notification and save the log

Appendix 4. License and Use Matrix

The license matrix links contract language to actual technical operations. It does not replace the contract, but rather serves as a tool for monitoring compliance and identifying gaps. For example, the right to "display materials within the service" should not be automatically translated into a permission for subsequent training of a universal model. Each matrix row should contain the text of the corresponding permission, its legal interpretation, limitations, validity period, and a supporting clause of the contract.

Objects with multiple copyright holders, audiovisual works, translations, photographs, and independent licenses for database data require special attention. A lawyer will coordinate the matrix with the product owner: the terms may depend on the specific architecture, territory, number of end users, and the availability of public scales. If there is any ambiguity, the company should obtain an addendum to the agreement or refrain from performing the disputed action; economic necessity does not compensate for a missing right.

Method of use	Status	Basis and limitation
Download and technical copy	Yes / No / Not clear	Contract clause or exception

Method of use	Status	Basis and limitation
Preprocessing and index	Yes / No / Not clear	Place, time, access
Training the base model	Yes / No / Not clear	Models and purpose
Fine-tuning of the new version	Yes / No / Not clear	Additional permissions
Downstream transmission	Yes / No / Not clear	Sublicense and territory

Appendix 5. Protocol for proof of legitimate access

Access to a web page must be proven by the circumstances that existed at the time of the relevant copying. The auditor records the exact URL, redirect source, subscription status, technical method of acquisition, and the text of the terms and conditions as of the relevant date. An archived copy may have evidentiary value, but must be marked as an external historical source with a disclaimer regarding completeness. A more recent screenshot does not automatically confirm the terms and conditions from a year ago.

For subscription publications, an additional check is made to ensure that the user's permissions extend to automated analysis, not just individual reading. For open websites, the presence of restrictions and opt-outs is checked. In the absence of reliable information, the conclusion is formulated as "authorized access not confirmed" rather than "theft proven." Actual authentication bypasses and the acquisition of material from a pirated archive are recorded separately and referred to a legal expert for qualification.

Artifact	What confirms	Main limit
HTTP log and response code	Technical route of receipt	Doesn't show copy rights
Subscription and its terms	Type of legal access	May limit automation
Snapshot of licensing terms	Fixed date mode	Requires authentication
Opt-out rules	Will of the copyright holder/service	Legal interpretation is needed
Hash of the downloaded file	Identity of the copy	Does not prove the legitimacy of the source

Appendix 6. Due diligence of the training data provider

Before purchasing a dataset, you should request a map of upstream sources, information on acquired rights, sublicense terms, a history of claims, and proof of compliance with copyright holder restrictions. A statement of compliance is useful as a contractual obligation, but without access to at least a sample of the source documents, it does not mitigate the risk. Particular attention should be paid to data collected by third-party robots, as the contractor may have lost or ignored redundancy signals during dataset preparation.

The contractor's due diligence should consider financial stability, the ability to preserve documents after contract termination, and the obligation to cooperate with claims. An independent sample verifies both small and large sources; assets with a high risk of illicit origin should be included. An indemnity clause without a provenance transfer procedure is incomplete. The internal assessment must specify which evidence the supplier provided itself and which was independently verified.

Request	Risk of non-provision	Possible control
List of upstream source categories	Unable to determine origin	Independent sample
Right of sublicense	Breaking the permission chain	Verification of contracts
Opt-out log	Unproven compliance with Article 4 of the DSM	Test on selected objects
Removal rules	Non-executable remediation	Technical walkthrough
Contractor's exit plan	Loss of evidence	Obligation to transfer the journal

Appendix 7. Public Reporting Checklist for Article 53(1)(d)

The public summary form must cover exactly the models and versions listed on the vendor's page. The auditor compares it with the mandatory AI Office template, published in 2025, and the internal provenance registry,

without substituting their own requirements for this minimum standard. Correct coverage of categories is important: public datasets, web collection, privately licensed datasets, user-generated and synthetic content, and significant subsequent updates. [3][17]

When certain information about older models is objectively unavailable, the document must include the reason and extent of the gap, if such an explanation is permitted by the relevant regime. Refusal to disclose trade secrets does not automatically exempt one from adhering to the template structure. The public form and the closed evidentiary statement must be comparable in identifiers, but do not have to contain the same depth. The auditor does not confirm the legitimacy of all works based on the fact that the summary appears complete.

Chapter	Security question	Red flag
Model Identity	Is the version and supplier indicated?	The summary refers to an unclear product
Source categories	Are all significant groups reflected?	Hidden all web crawl as "other"
Updates	Are dates and changes documented?	No delta after fine-tuning
Spaces	Are the limitations explained?	Assertion of absolute completeness without a manifesto

Appendix 8. Technical provenance graph and transformation control

The provenance graph must link the original object to each derived record: OCR extraction, normalization, block splitting, duplicate removal, embedding, and the training batch. When a file is processed, its hash changes, so storing a single control value without parent relations does not ensure traceability. An auditor must be able to select a random element in the training manifest and reconstruct the path to the original source and associated legal basis.

To protect personal data, the journal should not become a public catalog of individuals and works. A public report should be limited to aggregates and verifiable procedures, while private access to specific objects is provided as needed and under control. A revocation mechanism is particularly important: a change in the legal status of the original asset should trigger a review of all downstream derivatives. This can be tested on a synthetic control set without actual protected works.

Knot	Mandatory connections	Independent auditor test
Source object	URL/agreement/hash/date	Restore your legal card
Extracted text	Parent and Extractor	Check the connection persistence
Dedup group	All source objects	Don't lose the terms of different licenses
Training batch	Dataset version and assets	Random write backtrace
Model version	Used batches	Identify the area of risk involved

Appendix 9. Protocol for retraining and updating the model version

Further training can change applicable documents and public transparency. The protocol records the base model, legal provider, purpose of the additional corpus, its rights categories, the amount of new assets, excluded sources, memorization test results, and the date of the future summary update. It is necessary to legally assess whether the modification of a specific downstream participant transforms it into a general-purpose provider of the new model under the applicable AI Act criteria. Not every interface customization constitutes such a change of status. [2][3]

A mandatory component is the delta registry: which assets have been added, replaced, or removed since the previous version. An old audit report cannot simply be re-implemented if the data has changed significantly. The audit can be limited to new sections if there is proof that the rest of the chain has not changed; however, this proof must be positive and documented, not based on the developer's assurances.

Update field	Meaning	Solution
Base model ID	Legal entity and version	Check the applicability of duties
Added data	Categories and Provenance	Permission/Exception/Opt-out
Removed data	Basis and confirmation	Residual weight estimation

Update field	Meaning	Solution
Public summary	What and when was revised	Publish using template mode
Independent retest	Sample size and results	Admission, restrictions, escalation

Appendix 10. Testing the risk of reproduction of protected text

The protocol defines a fixed repertoire of works, copyright status, fragment selection, permissible queries, number of repetitions, and evaluation criteria before launching the experiment. It is prohibited to include long original text in the prompt and then consider its appearance in the response as evidence of memorization. Regular queries and targeted recovery attempts are performed separately, with separate statistics and qualitative analysis of the original elements.

Test logs store the model version, time, generation settings, filter parameters, exact query formulation, and match metrics. Publishing long, protected passages in a public report simply for the purpose of demonstrating the result is prohibited. A legal expert evaluates the materiality of the protected form, a technical expert evaluates reproducibility, and an editor evaluates the need for disclosure and the risk of secondary infringement. If the result is negative, the final test coverage is indicated, not a guarantee of non-reproducibility.

Criterion	Method	Acceptable result
Repertoire	Pre-fixed sample	Eliminate cherry-picking
Regular request	Context without a long original	Frequency of unintentional reproduction
Targeted inspection	Limited number of pre-defined queries	Risk of extraction during attack
Comparison	Uniqueness, length, identity	Qualitative legal analysis
Report	Intervals, limits, source logs	Reproducibility without content leakage

Appendix 11. Copyright Claim Registration Form

The claim form must ensure verification of the right, subject matter, and technical route, without requiring the applicant to prove impossible internal information. Information on authorship, a link to the original, the intended use, the nature of the damage, and the preferred method of response are accepted. The applicant is provided with the registration number, date of registration, the requested missing information, and the contact information of the responsible party. Unnecessary identification documents are not collected.

Within the organization, the claim is compared against the asset registry and model version. If the original object is not found, the result is marked "not found based on available evidence" rather than "never existed." If an active legal dispute is detected, the relevant logs are preserved, further dissemination of the proven replication is technically restricted, and a legal assessment of the permitted measures is conducted. The response contains the grounds and procedure for review, without disclosing third-party data.

Address field	To be filled in	Examination
Applicant	Author / Representative / CMO	Authority to exercise the stated rights
Work	Link, ID, proof	Distinguish between a work and a copy
Model behavior	Public withdrawal or collection	Binding to version and time
The requested measure	Answer, block, negotiate	Legal feasibility
Closing	Reasoned decision	Possibility of appeal

Appendix 12. Chain of Custody for Digital Materials

Each file submitted by the auditor receives an immutable identifier, a source hash, a timestamp, a description of the retrieval route, and a list of access rights. When copying, a derivative copy is created for analysis, while the original is stored separately. If the material was obtained via a screen recording or archived page, technical limitations of the format are recorded to prevent the derivative image from being misrepresented as the original document.

When evidence includes disputed works, access is limited to the necessary number of reviewers. Brief descriptions and hashes are used before publication of the report, unless a full fragment is required by legal procedure. For significant legal proceedings, a specialized lawyer is engaged to assess the admissibility of the evidence in the specific jurisdiction. A breach of the chain of custody must be disclosed, not concealed by a later restoration of the date.

Event	Mandatory entry	Responsibility
Receipt	Source, channel, date, hash	Custodian
Broadcast	Sender, recipient, target	Both participants
Transformation	Tool version, new hash	Technical expert
Publication	Volume of disclosure and base	Editor + lawyer
Archiving	Term, access, destruction	Evidence Controller

Appendix 13. Statistical sampling plan for a closed case

Before verification, a population map is created: the number of files, unique works, providers, languages, content types, and collection periods. Stratification takes into account both the share of the corresponding category and its risk. For example, a small licensed music catalog may require a separate stratum due to high potential impacts even with a small share of the overall corpus. Without a complete list of the original population, representativeness cannot be confirmed.

A random sample is generated using a reproducible method; the random source and seed are recorded. Targets selected from complaints are tested separately. The result is differentiated between lack of evidence, breach of contract, and established technical noncompliance. Statistical intervals are specified only when the assumptions of the corresponding model are met. The conclusion is not generalizable to the closed 40% of the corpus without additional data.

Parameter	Technical solution	Warning
Population	Deduplicated asset registry	Duplicate copies distort the share
Strata	Law, source, period, risk	Do not replace risk groups with average frequency
Seed	Documented generator	Cannot be changed after results
Gaps	A separate category	Don't consider "unknown" a match
Report	Share, interval, limitation	Don't translate a metadata defect into "piracy"

Appendix 14. Checking the agreement of two experts

For legally ambiguous objects, two trained assessors receive the same minimum information package but do not see each other's assessments. The instructions pre-distinguish between a valid license, a valid exemption, an unknown status, an established non-compliance, and an out-of-scope object. The primary labels and justifications are then compared. The discrepancy is reviewed by a third assessor, and the reasons for the correction are recorded in the methodology training log.

A formal attempt to increase consistency by lumping all complex cases into a vague category should be avoided. For each version of the rubric, examples of typical and borderline cases are retained, and a new court decision results in a dated review. If half of the cases were assessed under the old instructions and the other half under the new ones, the resulting score cannot be interpreted as a single test without a methodological caveat.

Indicator	Document	Management decision
Original labels	Separate forms of experts	Check the blind mode
Discrepancy matrix	Intersection of categories	Clarify ambiguous rules
Agreed conclusion	Reasoned decision	Fix dissent
New edition of the section	Version number and date	Do not overwrite previous markup

Appendix 15. Procedure for temporary suspension of use of the kit

A temporary suspension is not used as a pre-trial punishment, but rather as a measure to control potential legal and reputational risks. Triggers may include a confirmed pirated source, systematically ignored bans, lack of rights to a significant category, or the technical impossibility of localizing the distribution of the disputed material. The decision records specific versions, operations, and urgent measures to preserve evidence.

The shutdown must be targeted: blocking a specific source, a training batch, fine-tuning a process, or issuing a specific class of results, depending on the risk. The developer has the right to provide documents that dispel doubt; the auditor verifies them without a pre-determined outcome. Resumption is possible after a documented solution, elimination of the cause, and testing. Without this, simply renaming the corpus does not constitute remediation.

Trigger	Immediate action	Renewal condition
Pirate Archive	Block source and derived sets	New legal basis or exception
Broken opt-out	Stop future collection	Fixed parser and confirmed test
Provenance gap	Quarantine of unidentifiable part	Restoring the chain of evidence
Massive output risk	Feature/Version Limitation	Independent retest

Appendix 16. Verification of corrective actions and residual risk

After a fix is decided, not only is the deletion of the record from the table registry verified, but also the impact on copies, indexes, future training batches, and maintained models. For each fix, the technical scope, operator, and proof of result are specified. If the fix cannot be confirmed from the parameters of a previously trained model, residual uncertainty and additional safeguards are formulated at the delivery stage.

A negative test is useful: an artificially included control object should not be included in a future build after a ban has been imposed. This demonstrates the pipeline is working, but does not prove that all previous controversial elements have been resolved. The final conclusion notes what exactly was confirmed (for example, the blocking of a specific training bucket) and what remains inaccessible (for example, independent models from third-party developers).

Measure	Confirmation	What is not confirmed automatically
The original file has been deleted.	Hash and storage log	No derivative copies
Training batch is excluded	Manifest of the new version	Changing old scales
Added output filter	Regression tests	Impossibility of other workarounds
Notification sent to partner	Confirmation of receipt	Real removal from all partners

Appendix 17. Register of Independence and Conflicts of Interest

An independent report must disclose who is paying for the work, whether the parties have a consulting relationship, and who may influence the publication. The form includes the personal and institutional interests of the reviewers, previous employment with the reviewed provider, connections with copyright holders, and participation in disputed license negotiations. Not every connection automatically disqualifies an expert, but a significant circumstance requires a protective measure, replacement of the reviewer, or a reasoned limitation of the report.

The review agreement must separate the remuneration for the work from the positive outcome. Management has no right to remove inconvenient findings without a documented editorial decision. If independence cannot be ensured, the project may be framed as a client consultation rather than an independent review. The register must be updated regularly, especially when a new parallel assignment arises.

Conflict	Risk level	Reaction
Reward for a positive conclusion	Unacceptable	Change of contract or refusal
Recent employment with the person being audited	Increased	Disclosure and independent review

Conflict	Risk level	Reaction
Personal connection with the copyright holder	Contextual	Recusal by objects
Lack of access to evidence	Methodological	Limit report title and output

Appendix 18. Combined Map of Intellectual Property Rights and the GDPR

A map is required for corpora where a work contains personal data. The author's literary rights and the privacy of the person in the photograph are separate considerations. Even legally permitted copying of text for TDM does not address the question of whether it is permissible to process information about health, political views, or criminal prosecution. For each set, separate IP_basis and DP_basis fields should be assigned, as well as the result of a special constraint check. [16][20]

Public disclosure of personal information does not automatically entitle it to unrestricted use in a generative system. The purpose, minimization, required transparency, subject rights, and international transfers are reviewed. For sensitive client materials, exclusion from general training is recommended unless a separate, reliable legal solution exists. The auditor records both results separately to ensure that the absence of copyright risk does not obscure a data breach.

Object	IP check	Data protection check
Author's photo	Author/License/TDM	Persons depicted and the purpose
Judicial article	Publisher's Rights and Database	Information about the suspects
Biographical archive	Rights to text and images	Proportionality of processing
Client's letter	Document rights	Confidentiality and legality of transfer

Appendix 19. Case study: photographic catalog

The photographer grants the editorial office a license to post the images on the website. The aggregator downloads the images and sells the "open" set to the technical firm. The editorial office's license does not confirm the aggregator's right to commercialize the model or sublicense the original files. The audit is required to reconstruct the terms of the original agreement, verify the opt-out status at the time of collection, and separate the photographer's rights from any potential personal rights of the people depicted.

The technical test includes identifying duplicate files and modified formats, verifying authorship metadata, and comparing it with the downstream registry. The recommended solution until the gap is resolved is to quarantine the disputed category, request confirmation of the chain of rights, and document which versions of the model have already received the content. This case study is for educational purposes only and does not involve a real catalog, publisher, or model.

Case fact	Legal problem	Proof
Editorial license	Is training allowed?	Full text of the permit
Resale by an aggregator	Is there a sublicense?	Upstream contract
Metadata removed	Will we establish the author?	Photograph of the original and image identity
Several versions of the photo	Does opt-out apply to copies?	Perceptual imprint and provenance

Appendix 20. Case Study: Song Lyrics and a Generative Model

A training situation: the model regularly produces long, recognizable fragments of song lyrics in response to routine questions. Similar to the approach discussed in the GEMA case, the investigation examines reproducibility, the quality of literary form, the sources of instruction, and the actual control of the provider. However, the decision of the Munich court of first instance should not be automatically interpreted as the outcome of any new case; national law, licensing conditions, and technical details may differ. [11]

The procedure involves selecting independent testers, avoiding lengthy original prompts, documenting the model version, and accurately comparing protected fragments. The result separates potential playback in the model from the public output, as the legal classification of these operations differs. The public report uses descriptions of the results instead of multiple playbacks of the song lyrics themselves.

Step	Data	Criterion
Copyright holder	Text rights and powers	Scope of claim
Educational source	License or Exception	Applicability to a specific action
Output	Simple prompts and logs	Protected recognizable form
Remediation	Protection of the issue and the disputed corpus	Retest

Appendix 21. Case Study: Teaching Materials and Translation

The publisher produces a textbook and an authorized translation, and the university provides student digital access. The startup copies both languages, considering the university's subscription sufficient permission to teach its commercial model. The audit must separate the rights to the original, translation, and digital edition, verify the university's access conditions, and any applicable TDM exceptions. The presence of a research partner does not transform all of the startup's activities into research TDM under Article 3 of the DSM. [1]

A separate check is made to ensure that the model generates answers using long passages from the textbook, and not just uses learned concepts. Evidence preparation should prevent the distribution of a complete, unlicensed electronic version. The recommended measure is to restrict access to disputed files until a documentary verification is conducted, negotiate rights, make a separate decision on subsequent models, and update the public summary if necessary.

Knot	Clarification	Document
Source	Author and publisher rights	Publishing agreement
Translation	Independent rights	Translator's Agreement
University	Just access or training too?	Subscription terms
Startup	Scientific/commercial role status	Protocol of actual activities

Appendix 22. Conditional case: software code and open license

The company uses a public repository with code from various authors and licenses of various types. The developers refer to the entire set as open source, thus ignoring copyright notice requirements and specific terms of use. In practice, it's necessary to monitor the licenses of the repository, dependencies, individual files, and changes in terms between versions. Not every license requirement applies equally to training and distribution of the generated code, and this issue requires separate legal assessment.

The protocol is designed to identify exact and close matches of output fragments, but the automatic license detector does not provide a definitive conclusion without reviewing the facts. Cases where the model produces long, recognizable blocks of code with removed attribution notices should be recorded. In the absence of direct similarities, the output should not describe the license as irrelevant for upstream copying.

Examination	Entrance	Exit
Licenses	Repository files and versions	List of various conditions
Provenance	Source and hash	Connection with the training set
Output	Long similar sections	Technical Match Protocol
Documentation	Notice, attribution, distribution	Legal assessment of duties

Appendix 23. Case Study: News and Publishing Archives

The aggregator collects articles, photographs, and editorial summaries from news portals, then sells the dataset under general terms "for analysis." Legal assessment cannot be limited to the presence of socially significant facts.

The dataset may include the rights of journalists, photographers, press publishers, the producer of a commercial database, and the terms of access to a paid archive. It is necessary to separate the extraction of facts, the reproduction of text, and subsequent distribution to users. [6][7]

The auditor checks the volume of copied materials, the frequency of updates, the presence of opt-outs, aggregator rights, and the search results configuration. Particular attention is paid to the replacement of access to the original source: a service that returns a full paid article to the user is legally and economically distinct from a public fact search system. The report specifically notes the market consequences as a subject for future research, rather than as an automatically calculated loss for each publisher.

Object	Possible right	Examination
News text	Author and publisher	License and TDM
Photo	Photographer and other persons	Chain of rights
Archival database	Sui generis/conditions	Extraction scale
Full model response	Playback/Access	Behavior of a specific version

Appendix 24. Case Study: Museum Collection and Scientific Research

The museum provides digital materials to a research organization for the study of cataloguing techniques. At the same time, a commercial participant in the project proposes using the entire set in a universal image model. The analysis should verify which objects are protected, who manages the corresponding rights, whether special contractual conditions exist, and whether scientific research is separated from commercial use. The regime of Article 3 of the DSM does not automatically transfer to each partner. [1]

It is recommended to have two independent workspaces and separate data manifests. The scientific environment stores copies in a secure, established mode for verification of results; the commercial environment requires its own legal basis and approval decision. The auditor verifies that the contractor's access to research copies does not effectively equate to their free upload to the production infrastructure.

Parameter	Scientific environment	Commercial environment
Target	The research question being tested	Product development
Base	Checking the conditions of Article 3	License/Article 4 and other regulations
Access	Limited and protected	Separate legal decision
Exit	Publication of results	Checking model release rights

Appendix 25. Conditional case: court materials and commercial base

An independent research team is collecting published court decisions and wants to train a legal search system. A private aggregator is proposing to add paid annotations and classifiers without a separate license, arguing that court decisions are publicly available. The problem lies in the distinction between official texts, protected editorial materials, database rights, and the privacy of individuals mentioned in cases. [7][20]

The pilot must use only materials for which there is documented rights or a reliable basis, and exclude hidden client files. Criminal procedure materials must undergo a separate GDPR assessment; court publication does not always guarantee freedom of secondary learning. The public report must use hypothetical training data and a verified bibliography, rather than reproducing identifying details of a real case to illustrate the algorithm.

Layer	Legal problem	Measure
Official act	National Reuse Regime	Separate qualification
Publisher's Abstract	Copyright	Licensing check
Commercial search	Rights to the base and contract	API Terms
Personal information	GDPR and special regimes	Minimization and access

Appendix 26. API Provider as a Data Source for Further Training

The integrator receives responses from an external model via an API, collects them in its own database, and then uses them to train a competing product. It's necessary to analyze the API terms, rights to output materials, possible restrictions on training on the responses, the origin of protected fragments, and the rules for processing personal data. The statement that "the model's output is free because it's written by AI" is legally incomplete: the response may contain protected human elements or user data.

The auditor must establish which supplier actually generated the response, which prompt was used, who owns the rights to the source documents included in it, and what terms the integrator accepted. If the model accidentally reproduced a protected passage, further training on that response does not clear its legal history. It is recommended to prohibit the transfer of disputed output texts to the new training environment until the assessment.

Link	Document	Control
API terms	Version of the terms and conditions as of the date of application	Is re-training allowed?
Prompt	Source of user attachments	Privacy and IP
Output	Hash and log	Probable reproduction
Next model	Manifest of further training	Prevent inheritance of the defect

Appendix 27. Model for negotiating a license for a subject catalog

A negotiation matrix is proposed for the thematic catalog, in which the copyright holder grants not abstract "all rights," but specific permissions with verifiable parameters: types of work, educational objectives, time horizon, territory, reproduction rights and technical copies, attribution conditions, audit of the volume used, and claims processing procedures. Remuneration is negotiated separately from organizational administration costs.

The parties may define a mixed model: a fixed access fee and a variable component based on an agreed-upon methodology. However, the formula cannot be approved without a preliminary assessment of data quality and the contractual rights structure. A special issue is the fate of the trained model after termination: the contract must provide for realistic technical measures and not promise automatic erasure of any traces in the scales. For small authors, an accessible method for confirming the included works must be provided.

Subject of negotiations	Possible option	Risk of uncertainty
Repertoire	List of asset IDs and versions	Unable to confirm coverage
Permitted stages	Training / fine-tuning / evaluation	Different interpretations of "use"
Reward	Fixed / by volume / combined	Unverifiable calculation base
Termination	Ban on future collection and remediation procedure	The Unfulfilled Promise of Weight Removal

Appendix 28. Verification of the powers of a collective organization

A collective management organization may have a significant repertoire, but its powers must be defined in relation to specific uses and territories. A public performance agreement does not always cover recording, digital reproduction, or model training. For multiple rights holders, the distribution of payments, representation rules, and the possibility of conflicting claims are important. A single seal or letter cannot be considered to cover the entire scope of rights.

The auditor requests information about the representative's legal status, repertoire, delegated powers, geography, and the relevant category of use. A non-disclosure agreement may protect the personal details of the catalog, but the final report must clearly define the boundaries of the rights verified. Having a collective mechanism is useful for reducing costs, but it should not be presented as the only mandatory route for obtaining permission.

Document	Being checked	Potential gap
Charter/mandate	Powers of the organization	Category of rights not covered
Repertoire	Objects and exceptions	The authors have logged out.

Document	Being checked	Potential gap
Territory	Countries and cross-border agreements	The location of the action is not covered
Distribution	Principles and reporting	Unverifiable compensation

Appendix 29. Cross-border matrix of legal actions

For each stage, it is necessary to consider the technical location, the place of commercial use, the rights of the claimed copyright holders, and the likely jurisdiction of the dispute. It is not recommended to infer the legality of the entire transaction from the country of registration of a single contractor. Even if the copy was created outside the EU under a local exemption, further provision of the service within the Union may give rise to separate obligations under the AI Act and copyright issues. The precise conflict of laws is determined based on the specific transaction and facts.

The matrix helps avoid incompatible simplifications: the American fair use standard is not covered by Article 4 of the European DSM, the British consultation did not establish an automatic opt-out in 2026, and the Singaporean computational data analysis standard has its own terms. For an official opinion, a specialized lawyer in the relevant jurisdiction should be engaged, and the validity date of the regulatory map should be recorded.

Stage	Jurisdictional linkage	What is being assessed
Online collection	Actual place and right of source	Access, licenses, TDM
Education	Place of performance and applicable law	Copying and Exceptions
Release of GPAI in the EU	The role of the supplier and the Union market	Article 53 of the AI Act
Conclusion dispute	Court, territory and nature of rights	Reproduction/responsibility

Appendix 30. Audit of claims of “100% legal” training

The researcher receives a marketing claim from the provider that all training data is completely legal. This cannot be confirmed without checking the definition of "legal," the legal geography, the volume, and the period. It is recommended to compile a list of sub-assertions: the sources have a verified basis for access; exceptions were applied correctly; machine-readable objections were processed; licenses covered the required operations; the policy and summary comply with the AI Act, where applicable.

For each subassertion, the observed evidence, the inaccessible portion, and the residual risk are recorded. The report can confirm a specific procedure, such as "for the verified sample from source X, crawler bans were processed correctly during period Y," but it should not turn this into a blanket statement about billions of unknown objects. The inference formula must be verifiable and accessible to external critical review.

Claim	Examination	Correct wording
Complete legality	Access to the full source corpus?	Limit the scope and criteria
No piracy	Independent source control	"Not detected in the tested sample"
Respect opt-out	Negative Test Protocol	Completion percentage and gaps
Independent confirmation	Financial and editorial autonomy	Reveal significant connections

Appendix 31. Compliance Cost Assessment Matrix

An economic assessment should include both the model provider's costs for documentation, licensing, and auditing, as well as the authors' costs for reserving rights, verifying the inclusion of works, and filing claims. A comparison that includes one party's costs while ignoring the other is methodologically unstable. For small participants, it is important to highlight fixed costs: standard verification, consultations, catalog registration, contract revision, and staff training.[5][9]

Instead of the unproven prediction that "regulation costs tens of billions," a pilot project on a limited population with documented transaction costs is proposed, followed by a scenario analysis. The conclusion should indicate

sensitivity to data coverage, model scale, signal automation, and dispute frequency. Increased trust and reduced costs for future litigation are separately considered as potential benefits, but not guaranteed savings.

Expense item	Developer	Copyright holder
Register of assets	Conducting an inventory	Comparison of repertoire
Opt-out	Checking signals	Expression and control of prohibition
Permissions	Negotiations/payments	Representation/receipt of payments
Disputes	Evidence and examination	Legal appeal and monitoring

Appendix 32. Observatoire Public Error Correction Policy

If a published analytical report contains a factual inaccuracy, the organization must provide a public and traceable correction without silently replacing the historical version. The correction must indicate which passage was changed, why, when, and what sources the updated version is based on. For disputed materials about a specific company, an opportunity to submit factual comments should be provided without granting the defendant the right to pre-censor independent research.

The original version, the new document, the changelog, the publication date, and links to the correction are technically preserved. A significant error regarding the legal status of a license or the progress of a legal proceeding requires a clear note at the beginning of the text. A simple typo correction may be reflected in the general log. The protection of commercial secrets and personal data is assessed separately from the principle of transparency. Such a policy is essential for the credibility of independent research.

Event	Reaction	Public announcement
Factual error	Verification and documentary correction	Date and description
New court ruling	Addition without retouching the previous status	Updated context
Challenging by a party	Independent review	Summary and motivation
Dangerous disclosure	Urgent access restriction	Reasonably sufficient explanation

Appendix 33. ARGA draft responses for consultation with the European Commission

Recommended position on Technology and Copyright: Improving the transparency of training should enable copyright holders to verify the legality of use of a work, but should not create a public registry of commercially confidential contracts and technical vulnerabilities. A two-tier architecture is proposed: a public summary of significant categories and private sources according to an established template and a closed, auditable asset-level registry for legitimate regulatory and judicial proceedings. Interoperable provenance formats and the preservation of opt-out histories should be encouraged. [5]

Regarding reservation rights, a minimum machine-readable standard is proposed, accessible to small authors free of charge, with verifiable developer responses. Regarding licensing, transparent terms for distributing rights along the supply chain and accessible contractual structures for collective or individual remuneration. Regarding auditing, the independence of the auditor, publication of sample boundaries, and a protective procedure for accessing restricted data are proposed. This is a draft of a substantive position, not a confirmation of a submission on behalf of ARGA; the text must be adapted to the specific questions of the official form and agreed upon with management.

Topic of the consultation	Suggested answer	Need for evidence
Training transparency	Multi-level disclosure	Provider manifests and checks
Rights reservation	Free machine-readable signals	Compliance tests across platforms
Licensing	Traceable rights along the chain	Contractual sampling and compensation
Enforcement	Available dispute channel	A study of the costs of small authors

Appendix 34. Secure Data Room Protocol for the Auditor

To audit a closed corpus without publishing commercial catalogs, a secure environment can be established with access to a registry selection, relevant licenses, and technical processing logs. Access should be role-based, logged, time-limited, and with control over exported results. The existence of a clean room does not in itself constitute legal grounds for processing or transferring personal data: contracts, legality, necessity, and information categories are verified before access is granted.

An independent auditor must have the right to verify the completeness of the population; otherwise, the provider may only present convenient objects. Random queries according to an approved procedure, external verification of the sample, and proof of the immutability of the manifest are useful. All limitations preventing completeness verification should be disclosed in the final report. A closed environment with ten examples pre-selected by the provider should not be presented as an audit of a multi-billion dollar corpus.

Control	Implementation	Examination
Completeness of the registry	Versioned manifest	Comparison with pipeline
Role-based access	Separation of functions	Review logs
Export	Aggregates, permitted evidence	IP/GDPR Control
Independence	Random selection of auditor	Don't limit yourself to a convenient selection

Appendix 35. ARGA Ninety-Day Training Pilot

The pilot does not provide access to real closed systems without a separate agreement. During the first thirty days, legal criteria, a sample license, a synthetic reference corpus, and an opt-out processing methodology are developed. The next thirty days are devoted to setting up the source log, transformation graph, case documentation, and negative tests. The final thirty days are for an independent random check, a gap report, and methodology adjustments.

The plan must have stop gates: lack of copyright for the training material, a broken chain of access, or a conflict of interest that precludes independent conclusions. The pilot measures the feasibility of the methodology itself, not the market share of illegally trained models. Its success is determined by reproducible protocols and fair limitations; the number of tables and pages does not serve as a metric for success.

Period	Job	Completion Criteria
Days 1–30	Law, forms, educational data	Approved methodology
Days 31–60	Provenance pipeline, opt-out tests	Repeatable technical records
Days 61–90	Blind review and review	Limited Report
Stop gate	IP, GDPR, independence	Suspension until resolved

Appendix 36. Template of the final auditor's report

The report begins with the identification of the model being reviewed, the organization, the data version, the period, the jurisdictions, the client, and the role of the expert. This is followed by the regulatory assessment criteria, the level of access, the sampling methods, the sources of evidence, the identified nonconformities, the corrective actions, and the residual uncertainty. The terms "verified," "not identified," and "confirmed compliance" are not interchangeable. The latter applies only to a specific set of criteria and sufficient evidence.

The conclusion must not conceal any unreviewed portions of the corpus or any assurances provided only by the client. An audit trail must be provided for each material statement. Trade secrets allow for some information to be summarized in the public version, but do not remove the limitations themselves. The report must end with the date, the signatory's authority, the revision procedure, and the review status.

Conclusion section	Requirement	Invalid substitution
Scope	Accurate models and sources	"The whole company" without verification
Criteria	Norms and standards with versions	General Declaration of Ethics

Conclusion section	Requirement	Invalid substitution
Evidence	Sample, hashes, documents	Advertising assurances
Findings	Fact, right, uncertainty	Automatic accusation
Limitations	Untested parts and reasons	Default

Appendix 37. Procedure for requesting clarification from the model developer

The request must be specific and legally motivated. Instead of demanding the release of the entire training corpus without justification, questions are asked about the identity of the model, the origin of a specific category of sources, the applicable license or exception, the processing of reservations, and the availability of limited auditing. The inclusion of trade secrets requires a separate protection regime and a necessity assessment. The copyright holder's contact information and representative details are processed according to the purpose of the request.

The draft letter establishes the neutral status of the investigative review and explicitly excludes the assumption of a proven violation until the response is reviewed. If there is no response, the wording "information was not provided in response to our request within the specified period" is used, rather than "the company has admitted the violation." A repeat request requires a new material reason or the requested deadline, rather than automatic, endless correspondence.

Request block	Item	Expected response
Identifier	Model and version	Confirmation
Source type	Category/Supplier	Provenance summary
Legal basis	License/DSM	Reasoned confirmation
Opt-out	Implementation and tests	Procedures or restrictions

Appendix 38. Special regime for information on sanctions and international search

The training scenario concerns a model trained on investigative journalism, records from international databases, and court documents on alleged crimes. The legality of citing a public fact must not be confused with the right to mass copy an editorial catalog and the ability to further process information about a specific individual. Copyright verification is carried out in parallel with GDPR, potential privacy concerns, and the risk of defamation or misattribution. [7][20]

The recommended measure for ARGA is not to transfer client files to the shared training service, maintain a separate evidentiary history of the origin of each publication, and document corrections whenever a person's procedural status changes. If a model generates a false "wanted status," this is a matter of factual accuracy and privacy; it does not, by itself, prove copyright infringement. An independent determination keeps these categories separate.

Risk	A separate legal route	Control
Copying a commercial database	Copyright/base rights	License and Extraction
Criminal prosecution data	GDPR, Article 10 and national law	Access, purpose and accuracy
Incorrect face status	Correction and protection of individual rights	Traceable correction
Client confidentiality	Confidentiality	Not for general use.

Appendix 39. Risk checklist before publication of the report

Before submitting the report to an external audience, the editor checks the status of each legal act, the relevance of court decisions, the bibliography's connection to the text, and the absence of information that could reveal a classified source or someone else's work beyond the permitted scope. Particular attention is paid to the wording of judicial practice: a first-instance court decision does not constitute a final pan-European norm; a consultation is not considered an adopted law; a voluntary code does not receive mandatory certification status.

Technical verification includes document format, active hyperlinks, absence of font squares, correct numbering, table readability, presence of a logo, and prohibition of hidden personal metadata in file properties. Legal review and technical preflight are separate procedures; successfully opening a PDF does not confirm the accuracy of the regulatory analysis.

Control	Responsible	Admission conditions
Standards and dates	Legal expert	Confirmed by primary sources
Factual statements	Analyst	Verifiable link and disclaimer
Quotes and IP	Editor	No other person's rights have been violated
Word/PDF	Technical editor	Opens without squares
ARGA approval	Management	Separate approval

Appendix 40. Matrix for monitoring changes after October 2026

The law of model training is still evolving. Therefore, the final publication must be accompanied by a list of events requiring review: the adoption of new mandatory regulations, the finalization of the Commission's consultations, changes to the official GPAI template, significant court decisions, and new machine-readable signal standards. Monitoring does not automatically change conclusions: each trigger is assessed for its legal relevance for a specific section.

Updates must retain the original version of the research and include a dated changelog. If a decision on a specific dispute is received, the editor verifies the authority, the appellability of the matter, the subject matter, and the legal system. If a draft guideline is published, its status is indicated separately. This disciplined process prevents the common mistake of using a news headline as if it were a final legal amendment.

Future event	Verifiable fact	Editor's decision
Closing of the consultation	Is there a final document?	Update your status, don't make up a solution
Judicial act	Instance, subject, appeal	Revise a specific thesis
New version of the AI Act	Date of application and coverage	Correct the calendar
Opt-out technical standard	Status and implementation	Update methodological tests

Appendix 41. Protocol for matching photographic rights with file versions

The protocol's goal is to establish the origin of a specific image without falsely inferring that a similar file necessarily belongs to the original user. The investigator records the unique file ID, size, format, checksum of the original copy, the found URL, date and method of acquisition, possible EXIF fields, version number, crop history, and author signature information. Data received from the provider is separated from independently verified data. When attempting to compare the original RAW file, JPEG export, and advertising resize, similarity parameters and alternative explanations are recorded. Perceptual similarity is a reason for a documentary request, but is not a substitute for attribution.

If the photographer's rights are claimed, the photo shoot commission, exclusive or non-exclusive license, agency authority, limits on sharing with third parties, methods of use in training, and further training activities are checked. The "source unknown" field cannot be replaced by the assumption of a free license. If the author's attribution is disputed, the auditor requests written confirmation and temporarily marks the object "requires verification" without declaring one of the parties in violation. A separate mark applies to portraits containing personal data: the presence of copyright permission does not resolve GDPR issues.

The protocol summary contains three types of conclusions: "the chain of rights is confirmed within the examined scope," "the chain is not confirmed due to a specific missing document," or "contradictory evidence was identified." The third condition should not be forced to be reduced to the first or second. A photograph is not reproduced in its entirety in a public report without independent justification for its reproduction. [6][20]

Appendix 42. Test of machine-readable rights reservation when changing URLs

The test environment simulates the sequence "original page - 301/302 redirect - CDN - archive copy - image file." For each stage, the URL, access rules, server responses, robot settings, retrieval time, and saved rights reservation signal are pre-recorded. If the copyright holder has updated the request, a check is made to see if the crawler has received it during the next crawl and whether a version of the solution is saved. A missing signal cannot be considered permission without confirming other necessary conditions for legitimate access.

The test includes three predefined checks. First, the disallowance is present on the primary URL and persists after a redirect. Second, when a rule is updated mid-session, the system re-evaluates the source's eligibility for inclusion. Third, the system rejects a mirror site as an independent, newly-owned resource simply because the domain string has changed. For each test, the expected and actual results are listed; failures are classified by origin (crawler, rights handler, data provider, manual exclusion).

The legal commentary should remind that the technical specification of signals and the legal meaning of their use are separate issues. Machine readability is relevant for online content under Article 4(3) of the DSM, but conflicts between specific signals, parties, and jurisdictions are subject to separate assessment. The test set should only contain data to which the researcher has the right to access and reproduce. [1][21]

Appendix 43. Working example of checking a mixed archive catalog

A hypothetical organization acquires access to a catalog of 20,000 items: 12,000 newspaper articles, 4,000 photographs, 2,000 audio recordings, and 2,000 editorial memos. These numbers are for educational purposes and do not represent a real company. The archival agreement permits reading and internal searches by employees, but makes no provision for retrieval for commercial training. The team mistakenly assigns the entire catalog the "licensed" status. The auditor splits the dataset into a four-type stratified sample, highlighting rare categories, and records permission uncertainty separately for each type. Problems with photographs and audio files are not masked by the high quality of the documentation for the articles.

The test begins with 40 random objects of each type, 160 in total, and does not automatically generalize the results to the entire population without weights, selection rules, and clustering considerations. For each object, the access date, essential rights, license, reservation, and authorization confirmation are entered. In the hypothetical group of 40 audio recordings, 9 have an incomplete chain of related rights. This is a figure of 22.5% for the subsample examined, not 22.5% of legally proven infringements, much less an estimate for all songs on the internet.

The recommendation is to suspend the inclusion of unapproved data types in the new training cycle until clarification is received, retain the audit documents, and request contractual confirmation from the supplier. If it is determined that the data is acceptable for research purposes under the established conditions, the conclusion will be reconsidered. The report should clarify that the lack of proof of licensing does not, in itself, preclude the existence of a legal exception.

Appendix 44. Comparison of GPAI and Public Summary of Learning Content Versions

For the notional developer's model, a version table is compiled: date of placing on the EU market, original base model, date of significant additional training, categories of newly used data, newly published summary, and responsible person. The check is conducted at the level of the identifiable product, not the abstract brand. If several versions have identical summary content, the admissibility of such disclosure is assessed based on the Commission's official explanations; if additional training adds new source categories, the need for an update must be assessed separately. [2][3][17]

Example: The base model M1 was trained on three categories; version M1.1 received additional legal corpus from a commercial vendor. The auditor does not require unconditional disclosure of all client agreements in a public document, but must ensure that the new class of sources is adequately reflected and that protected evidence of origin is available. The log records the due date of the next review and the reasons for the early update. The Commission's official FAQ describes periodic updates with a six-month target, or earlier if materially significant

new data is available; this target cannot be turned into a standalone provision for the automatic expiration of all copyright permissions.

The control result is presented as "summary published and assessed against template," "summary requires clarification," or "verification impossible due to insufficient data." A public summary does not confirm the validity of each corpus element, and the continuity of model versions must be reproducible through documentation.

Appendix 45. Test of author's texts without reproducing the full work

The auditor creates a small, legally accessible control sample of creative texts with predetermined usage criteria for testing. The protocol records the title, version, source, license type, inclusion reasons, originality indicators, and a prohibition on publishing the full content in the report. Before querying the model, the limits of acceptable prompts, the number of iterations, temperature and filter settings, and the time interval are set. The protocol distinguishes between routine user questions, attempts to reconstruct a rare fragment, and tasks involving reference data provided by the researcher to the model.

The result is assessed not only by symbolic similarity. The fact of matching a commonly used phrase is legally distinct from the reproduction of a long, individualized structure, and similarity of ideas or genre clichés does not constitute borrowing a protected form. Original elements, fragment length, the possibility of independent similarity, and the role of the provided input data are compared. Experts maintain the differences between technical and legal assessments rather than averaging them into a single, unexplained score.

A secure form is used for publication: statistics, an anonymized identifier, a short description of the features, interval metrics, and a reproducibility method with minimal disclosure. Even identified reproducibility in the test does not prove the legal status of the original training without additional materials. A separate legal assessment is possible after establishing specific actions and applicable exceptions. [6][11]

Appendix 46. Digital Music Library Agreement Card

For a hypothetical library of 800 phonograms, the auditor creates four linked tables: work, performers, master recording, and machine-readable licenses. A single composition title does not represent all authorized parties. Each recording is assigned an internal ID, information about the author of the lyrics and music, a phonogram identifier, performer, geography, term, and contractual basis. The presence of a public performance license does not automatically populate the "use for further training" field. [6][19]

When an intermediary document is discovered, not only its authenticity but also its mandate is verified. Issues include territorial jurisdiction, types of works, method of commercial use, rights of transfer to a technology contractor, and the catalog correction procedure. The absence of one part of the rights chain is recorded as a specific evidentiary problem. To avoid false claims, the result may contain the note "there is no document confirming a specific right in the available package" rather than the unconditional "right does not exist."

The model's output is validated separately for verbatim lyrics, recognizable fragments of the recording, overall stylistic similarity, and voice similarity. Each category uses its own test and legal analysis. The method preserves the date, model, specific queries, and conditions; the output is not transferred to other versions of the model without appropriate evidence.

Appendix 47. Example of license reconstruction through a chain of intermediaries

The study involves author A, publisher B, aggregator C, kit provider D, and developer E. Each link must demonstrate specific rights and uses, not just the fact that the previous participant was paid. Specifically, B's right to host a work on the platform does not necessarily encompass C's right to extract it en masse, and C-D's data provision agreement does not create a greater scope of exclusive rights than C had. Subcontracting and sublicensing are examined separately. Rights transfer dates are compared with the actual upload date.

For each link, the contractual clause, territory, term, object, right of further transfer, and supporting document are specified. The verifier specifies the level of availability of evidence: original, certified copy, extract, written statement, or unverified entry in the CRM. A missing contract is not filled in with a generic phrase about a

"partner copyright catalog." If the information contains a trade secret, access to the full document can be restricted to a secure data room with verification of the relevant provisions.

The summary contains a rights path with gaps identified, and, if necessary, an alternative legal basis, such as the applicability of the established TDM exception. It cannot be automatically assumed that all training is only possible by contract, just as it cannot be assumed that the exception always applies to all public content. [1][6]

Appendix 48. Instructions for a secure response to a copyright holder's request

An incoming claim may contain the author's name, a description of the work, a link to the disputed source, the basis for the copyright, and a request for exclusion. The Secretariat records the date, channel, sender, and unique ID, without distributing personal information in the public work chat. The first step is to verify whether the object belongs to the model and version being reviewed. The applicant is asked to provide only those documents necessary to identify the object and its rights, without unreasonable demands to hand over a passport or all commercial contracts.

The team then separately evaluates the prospects of stopping future collection, the legal position regarding existing copies, and the potential for replication in model outputs. These issues cannot be reduced to a single "delete" button. Any claim of complete deletion from the scale requires technical evidence, which many systems lack; therefore, the promise is formulated strictly based on actual actions. If the company did not have access to the source code in the claimed corpus, the response must confirm the scope of the verification performed and its limitations.

Finally, the applicant receives a clear response listing the objects examined, the measures taken, and any outstanding issues. The log records the review of the decision based on new documents and the right to appeal to the competent authority. If the claim concerns both the authorship of a photograph and a personal image, the claim is referred to both competent authorities, without confusing the applicable legal grounds. [1][20]

Appendix 49. Control over the publication of the expert opinion of Observatoire ARGA

Before publication of the report, an editor not involved in the fact-finding process is appointed to check the text for compliance with the internal evidence ledger. Each legally significant statement is assigned a document ID, review date, jurisdiction, norm status, and review owner. Potential conflicts of interest are considered, including paid consulting by one of the parties, research funding, employee involvement in a related dispute, and previous public assessments of a specific company. A conflict does not always preclude publication, but should lead to a pre-arranged independent resolution of the review or recusal issue.

The confirmed fact, legal assessment, analytical risk, hypothetical scenario, and hypothesis are separately identified in the text. If the audited company has not responded to the request, the statement "it admitted the violation" is not permitted. Accordingly, representatives' responses are considered, indicating their evidentiary status and possible verification. It is prohibited to conceal facts unfavorable to the preliminary hypothesis or retroactively change the sampling rules for the sake of a desired result.

The public version does not fully disclose other people's protected works, documents containing trade secrets, personal information without justification, or secret access identifiers. The appendix for internal reviewers may contain protected evidence. Both versions must have a common numbering of conclusions and a correspondence table for subsequent verification.

Appendix 50. Control of the cross-border judicial scenario

A hypothetical developer is located in country X, trains a model with the participation of a data provider from Y, and makes the product available in the EU. The legal analysis begins with a map of actions and territories, rather than automatically transferring any national judicial position to the entire chain. It is necessary to distinguish the law governing the relevant exclusive right and potential tort from the law governing the data supply contract and the procedural rules governing access to evidence. Having a presence in the EU does not mean that every claim for copying abroad is resolved by the same substantive rules.

US court decisions on fair use in individual circumstances cannot be mechanically used as a substitute for an analysis of Articles 3 and 4 of the EU DSM. It is equally misleading to present the German first-instance decision on memorization as a universal answer to the legality of copying all books, songs, and images. For each position, the lawyer identifies the jurisdiction, the stage of the decision, the factual situation, and the limits of its precedential value. [1][11][12][13]

The practical summary includes a map of potential competent authorities, anticipated types of claims, evidence security risks, confidentiality rules, a plan for inquiries with counterparties, and a warning about necessary local consultations. This map is an analytical tool, not a legal opinion on a specific case without a factual package.

Appendix 51. Conditional calculation of royalties without arbitrary presumption of damage

The initial values of the example are arbitrary: the catalog contains 2,000 works, of which use for the agreed-upon product is agreed upon for 1,500. For the remaining 500, there is no confirmed contractual basis, but this does not yet establish infringement due to possible exceptions and other legal grounds. In the negotiation model, the parties consider the price of access for a certain period to a specific type of use. If the agreed-upon rate were €8 per item, the illustrative amount for 500 items would be €4,000, but this calculation does not determine the actual amount of damages or the market fairness of the rate.

A full financial assessment requires establishing the relevant licensed market, comparable transactions, the volume of actual usage, the substitutability of the product, the possibility of distribution among different rights holders, and the termination conditions. A notional rate can serve as a starting point for negotiations, but it does not prove that the court should award this exact compensation. A single work is sometimes used across multiple rights, and an unjustified multiple assessment leads to an inflated amount.

The document recommends maintaining three separate figures: measured actual revenue, observed content volume, and the estimated negotiation fee. These should not be combined as if they were a single variable. When published, the calculation method and the impact of key assumptions on the outcome should be disclosed; sensitive commercial information may be shown in a range or in a closed appendix.

Appendix 52. Register of versions and critical dates of legal review

For each published study, a master register is created: the report version, the date of the legislative review, the date of the official FAQ review, the version of the model examined, the research dataset, the relevant court decisions, and any outstanding issues. A change to the regulatory text generates an impact assessment entry for the sections, rather than automatically overwriting the historical version of the study. The European Commission's Technology and Copyright consultation for the date of 23 October 2026 is still open until 3 November 2026; it cannot be presented as a final amendment to the directives. [5]

The register provides a source reference, exact date, status (law/act/court decision/official clarification/consultation/research proposal), and the responsible editor's action for each significant event. Special verification is required when publishing appeals, updating the GPAI template, or adopting new technical specifications for rights reservation. An archived copy of the research version preserves previously reached conclusions and allows one to identify which ones are outdated.

The final step is reviewing the publication package. The Word and PDF versions receive identical details, logo, and version name; sources contain active hyperlinks; the content references the correct pages; and all quantitative examples are designated as case studies. Signatures or institutional approval are not added without a direct decision from management. The final editorial review separates the technical preflight, legal review, and publication decision.

Appendix 53. Table of TDM and Model Output Delimitation

The auditor creates two independent records for a single work. The first reflects the life cycle of the input: the date of legal access, the creation of temporary copies, the applicable Article 3 or 4 of the DSM, the presence of reservations, and the result of the audit. The second reflects the behavior of the output: whether reference requests

were made, whether significant original elements were reproduced, and what exceptions and contractual permissions may be relevant. One action does not automatically justify the other.[1][6]

A practical example: an archived text was legitimately used for research analysis, but the model then produced a lengthy, characteristic fragment without the user providing the original. A reference to the research TDM alone does not complete the analysis of the second action. The opposite situation: a similar general retelling of facts in the model's response does not prove that the training corpus contained a copyrighted article. The journal specifies separate proof standards and does not allow a single "copyright cleared" checkbox.

Stage	Document	Independent question
Gaining access	Agreement or conditions of legal access	Is it possible to explore the object?
Creating a training copy	Actual transactions and exclusion	Is playback acceptable for TDM?
Transfer to partner	License/terms of engagement	Is there a right of further use?
Model issuance	Protocol, matches, login mode	Is there a separate use of the protected form?

Appendix 54. Matrix of copyright holder objections by content category

Copyright holder applications are categorized into text, image, audio, video, code, and databases. For each category, a mandatory minimum is created: object identifier, initial receipt date, applicant, right type, authority verification level, specific required result, and contact information of the responsible party. Requesting unnecessary biometric or passport information to verify rights to a work should not be done without specific justification. If the application is submitted by an agent, their credentials are verified.

A separate status is available for each object: registered; clarification required; basis verified; temporarily excluded from new collection; measure implemented; refusal with justification; open for review. These statuses are not equivalent to judicial recognition of rights. The technical team confirms the actual change to the dataset or blocking, the legal team confirms the purpose of the action, and the editor confirms the accuracy of the response to the applicant. An external response should not promise unconditional "erasure from the scales" if such a measure has not been technically implemented and proven.

A quarterly control sample of complaints allows us to assess the average response time, the proportion of corrected classification errors, and the frequency of calls to a single source. These statistics should include the number of reasons, not just percentages, to prevent a small sample size from appearing like a consistent trend. If a widespread error is detected, the escalation mechanism targets the source category, not just the individual complaint.

Appendix 55. Scheme of access of an independent auditor to a closed building

The data room standard is defined by an agreement on objectives, access limits, responsibility for information, and the procedure for recording actions. The auditor specifies the set of audits in advance, but may reserve the right to randomly select them according to a pre-agreed methodology. The client must not substitute the sample for purely convenient examples. Document reviews, their extracts, and the audit itself are recorded in a log; full files are not uploaded to open-source tools or used for third-party models.

Access control has three layers: a technical specialist checks versioning and origin; a lawyer evaluates relevant contractual fragments; and an external reviewer receives an anonymized evidence ledger and reproducible test parameters. For each layer, the need for visibility of the protected work itself is justified. At the end of the session, a record is kept of what information was retained, what was destroyed, what must be retained due to a dispute, and who confirmed the completion of the procedures.

It is necessary to describe in advance how the audit will proceed if a key document is refused. Acceptable conclusions include "verification is limited," "the supplier provided only an assurance," and "there is no evidence for this conclusion." Denying access does not replace proof of a violation. A closed access policy may protect commercial confidentiality, but it does not justify a positive independent opinion in the actual absence of the materials being audited.

Appendix 56. Protocol for stratified sampling of sources

Before extracting a sample, the auditor determines the population, time period, source categories, known clusters, key hypotheses, and expected error rates. Technically convenient files pre-selected by the vendor for presentation do not constitute a random sample of the corpus. For a mixed catalog, stratification is acceptable: books, news, images, music, code, specialized archives. Sample size may vary by category, but the final assessment should take into account weights and inclusion probabilities.

A case study: 10,000 elements are divided into 8,000 texts and 2,000 photographs. If the auditor examines 100 from each group and finds 10 uncertain legal grounds for the texts and 20 for the photographs, combining 30/200 = 15% into an unweighted estimate of the entire corpus is incorrect. With random representative samples, the crude weighted estimate would be 0.8 x 10% + 0.2 x 20% = 12%, ignoring confidence intervals and correlation. However, an "uncertain ground" does not equal a proven violation.

The result must separately report the number and proportion of unverified objects, the data availability limits, the exclusion criteria, and the selection method. If randomness is not ensured, the conclusion is called a target sample check, not a statistical evaluation of the entire corpus. Distinguishing between the actual lack of a license and the probability of a valid legal exemption is especially important in summary tables.

Appendix 57. Regulations for checking the version of a document during an EU consultation

For each public consultation, the source, URL, publication date, completion date, list of official materials, and working version of the expert position are recorded. Report No. 23 analyzes the targeted consultation Technology and Copyright dated September 29, 2026; as of October 22, the deadline for submitting a position is November 3, 2026. Until the official completion and subsequent regulatory effect, participant proposals do not constitute adopted legislation. [5]

Before submission, ARGA's mandate for the position, the approved organization name, the language, the non-disclosure of client information, the evidentiary validity of the statistics, and the correctness of the reference to a specific survey question are verified. Hypothetical cases are designated as hypothetical, and actual statistics are provided only with a verified collection method. A distinction should be made between legal proposals, comments on technical implementation, and requests for additional research; mixing them creates the impression in the recipient that the organization is demanding an immediate change to the law to address a single technical issue.

Submission monitoring includes checking the updated consultation text, confirming registration, and archiving the submitted position itself in an immutable version. An internal report records the actual status of "prepared," "sent," or "confirmed." It is prohibited to declare the submission or acceptance of a position without the corresponding document.

Appendix 58. Catalog of Prohibitions on the Use of Audit Methodology for Other Purposes

A list of limitations is published along with the methodology. It does not establish photo authorship based on a single perceptual hash, does not prove the origin of a specific text from the training corpus based solely on response similarity, does not grant a universal training license under Section 53 of the AI Act, and does not confirm the legitimacy of each transaction based on the overall vendor score. Tests on the outputs of a single model are not automatically carried over to all future updates.

Specifically, incorrect wording in advertising materials is prohibited: "certified by Observatoire ARGA" without a valid procedure and established standard, "100% copyright compliant" based on a small sample, "damage established" without an economic analysis, and "the court upheld permission for any training" when citing a decision on a specific dispute. For each public indicator, the subject and scope of testing, the date, persons with conflicts of interest, and the auditor's level of access must be specified.

Violation of these conditions will result in mandatory editorial review of the published material and, if necessary, correction. If a counterparty continues to use a research report as evidence of authority it does not grant,

Observatoire has the right to publicly clarify its true scope and seek a legal review of the situation. This measure protects both the research reputation and the interests of the authors.

Appendix 59. Checking the compatibility of licenses and sublicenses for code

The verification process begins with identifying the repository version, commit hash, module paths, dependencies, license text, and checkout date. Priority is given to the primary licenses and terms of specific code, rather than the aggregated "MIT license" indicator on the repository card. Files with different licenses, fragments of third-party libraries, and changes to copyright notices are taken into account separately. [6]

For a training product, the auditor creates a matrix: object; copyright holder; license type; right to copy; required notices; possible distribution restrictions; right to transfer the material to a contractor; potential significant reproduction in the output code. If the terms require preserving the license text, the removal of the notice during data preparation is recorded as a specific risk. However, this cannot automatically infer the method by which the model will subsequently produce the code.

The report contains separate categories: contractual uncertainty, a license with special conditions, possible reproduction of a protected fragment, and information security of the generated code. The last category requires a different standard of review. The auditor notes cases where applicable law may allow for an exception to the TDM, and does not assert that all open source publications necessarily require a separate agreement.

Appendix 60. Comparative Opinion Form on US and EU Judicial Practice

For each case, the following information is recorded: forum, jurisdiction, date, procedural stage, subject matter of copying, proven facts, issue for decision, and scope of legal conclusion. This includes the decision on the partial fair use issue in Bartz v. Anthropic of June 23, 2025, the decision on the motions in Kadrey v. Meta of June 25, 2025, and the separate final order of July 20, 2026, approving the settlement in Bartz. They represent different stages and consequences and therefore should not be combined in the registry under a single status "case decided in favor of AI."[12][13][22][23]

A separate line is provided for the Munich court's decision in the GEMA v. OpenAI case from November 2025, with its first-instance status, memorization/output subject matter, and uncertainty of subsequent review. A national judgment does not become a pan-European norm simply because it concerns a well-known developer. A comparison with the DSM Directive is made, highlighting the differences in substantive tests and national regimes. [11]

The final matrix does not contain a single column for "AI win/rights holder win." Instead, it identifies the specific legal assertion supporting the decision, the fact that distinguishes it from the case under consideration, and any outstanding issues. This format prevents the use of incomplete press releases in lieu of legal reasoning.

Appendix 61. Checking the publishing set after changing the license

In the training scenario, the developer uploaded editorial materials in January; the license terms changed in March; and a new version of the model is being prepared in September. The first task is to establish the legal basis for the January changes and the scope of subsequent changes. It is impossible to automatically conclude that any new ban retroactively cancels all previous actions, or that the January permit continues to cover any new collection and commercial use. It is important to preserve historical versions of the terms, approval confirmations, and actual access times.

The registry must show which files are already included in the original training corpus, which were included in new cycles, and which were retrieved for the first time after the conditions changed. Issues of further use and the technical memory of the model are considered separately. If an opt-out is involved, Article 4 of the DSM, the proper form, the time of the reservation, and applicable national characteristics are checked. A decision is not made solely by comparing dates without analyzing the contract and the nature of the actions. [1]

The final table suggests measures: stopping new collections, maintaining evidence of past legality, negotiating an additional license, correcting the public summary if necessary, and limited testing of output playback. No step is stated as a general legal requirement without an independent justification.

Appendix 62. Protocol for verification of personal data in a licensed archive

The copyright license for a journalistic archive and the legal basis for processing personal data are two independent matters. The conditional catalog contains biographical information, photographs, references to legal proceedings, and the names of third parties. The auditor identifies personal data, possible special categories, and information about criminal offenses, as well as the purpose of further training. However, any reference to a court should not be considered automatically as a proven criminal record. For data falling under Articles 9 and 10 of the GDPR, a separate analysis of special conditions is required. [20]

Each field is assessed for minimization, accuracy, relevance, risk of combination with other sources, legal basis, international transfer restrictions, and processing of subject requests. When using data in an AI model, the possibility of later reproduction of a false or outdated status of an individual is checked. The copyright agreement with the publisher does not relieve the obligation to correct erroneous personal information or properly process requests.

The final solution may recommend excluding certain fields from the new training, pseudonymization in the audit repository, strengthening quality control, and maintaining evidence in a closed mode. A public report on such a case should demonstrate a typical problem without disclosing the personal details of a specific client or case.

Appendix 63. Protocol of internal review before approval

The reviewer receives the final Word and PDF versions of the report, along with a list of references. They check that the references correspond to the current version of the regulations as of the report date, that original decisions are not presented as final, that the calculations are confirmed, that the sample sizes are clearly stated, and that copyright terms are used consistently across all sections. Each comment is categorized as "legal accuracy," "evidence," "method," "editing," "layout," and has a priority.

Changes are made by noting the affected section and briefly justifying it. If a correction introduces a new source, it should be linked to the text, and if it removes a significant conclusion, it is necessary to check for any remaining references in the summary and practical appendices. Upon completion, a new PDF export is performed, the table of contents is recalculated, and a visual inspection of the title page, tables, headers and footers, and the end of the bibliography is performed. A successful technical review does not automatically imply legal approval.

The reviewer indicates the status: suitable for presentation to management; clarification needed; significant substantive risks; publication not yet recommended. The date and signature of the responsible manager cannot be added on their behalf without separate confirmation. Internal use of the project and official publication differ.

Appendix 64. Final editorial register of unresolved issues

By the time the study is published, questions for which the existing database does not provide a definitive answer are identified: the classification of certain types of learning in national law, the legal effect of specific machine reservations, the completeness of transactional licenses, the availability of evidence for closed corpora, the legal status of individual generation results, and the consequences of additional learning. For each question, significant alternative interpretations, the necessary new evidence, and the event requiring reconsideration are identified.

An example entry: "Is the reseller's right to license the archive for commercial training confirmed?"; status "not documented"; required material "authorization agreement"; responsible person "legal group"; critical deadline "until next upload cycle." Another entry: "Is there a risk of fragment reproduction when running the version V model?"; status "not assessed"; required material "test protocol without reference input"; responsible person "technical expert". These questions are not lumped together into one impersonal general assessment.

The register is published in abbreviated form along with the research report to clearly define the scope of the conclusion. It serves as a tool for preparing a new version of the report, but does not indicate the completion of technical tests or the receipt of documents that did not actually occur. When an item is closed, the date, evidence, and reason for the status change are retained.

Sources and normative and methodological bases

The regulatory and methodological framework review date is October 23, 2026. The numbers in square brackets within the report correspond to the following documents. Regulatory acts, court decisions, draft laws, and voluntary recommendations vary in legal status.

[1] Directive (EU) 2019/790, Articles 3 and 4: text and data mining exemptions. The European Parliament and the Council; current EU law.Open the original source

[2] Regulation (EU) 2024/1689, last amended on 27 July 2026: Articles 53 and 111.EUR-Lex; AI Act; current EU law.Open the original source

[3] GPAI Public Learning Content Summary Template FAQ. European Commission; official clarifications.Open the original source

[4] Questions and answers on GPAI Code of Practice. European Commission; voluntary instrument and official interpretations.Open the original source

[5] Technology and Copyright: Targeted consultation dated September 29, 2026. European Commission; consultation open until 3 November 2026.Open the original source

[6] Directive 2001/29/EC on copyright in the information society.EUR-Lex; right of reproduction and communication to the public.Open the original source

[7] Directive 96/9/EC on the legal protection of databases.EUR-Lex; sui generis right of the database manufacturer.Open the original source

[8] Guidelines on obligations for General-Purpose AI providers. European Commission; Official Interpretation of the Scope of the GPAI.Open the original source

[9] Report on Copyright and Artificial Intelligence dated March 18, 2026.UK Government; official consultation report.Open the original source

[10] US Copyright Office, Copyright and Artificial Intelligence, Part 3 (pre-publication), May 2025. Copyright Office; an analytical report, not a law.Open the original source

[11] LG München I, GEMA v OpenAI, 42 O 14139/24, court report of 11 November 2025. Bayerisches Staatsministerium der Justiz; decision of the first instance, not final as of the date indicated.Open the original source

[12] Bartz v. Anthropic PBC, Order on Fair Use, June 23, 2025.US District Court, ND California; for a separate decision, see [22] for a later disposition of the case.Open the original source

[13] Kadrey v Meta Platforms, 25 June 2025: WIPO Lex case summary.WIPO Lex; US case law and explanation of the scope of the decision.Open the original source

[14] WIPO Learning Machines: An Introduction to AI and IP for SMEs, 2026.WIPO; comparative methodological study.Open the original source

[15] Singapore Copyright Act 2021, sections 243–244. Singapore Statutes Online; computational data analysis.Open the original source

[16] EDPB Opinion 28/2024 on AI models and personal data. European Data Protection Board; 18 December 2024.Open the original source

[17] Explanatory Notice and Template for Public Summary of Training Content, July 24, 2025. European Commission; template and explanations.Open the original source

[18] General-Purpose AI Code of Practice. European Commission; voluntary code.Open the original source

[19] TRIPS Agreement, Article 13: limitations and exceptions. World Trade Organization; international three-step test.Open the original source

[20] GDPR, Regulation (EU) 2016/679.EUR-Lex; lawfulness and accountability of personal data processing.Open the original source

[21] Code de la proprieté intellectuelle, article L122-5-3.Légifrance; French TDM exceptions.Open the original source

[22] Bartz v. Anthropic PBC, Order Granting Final Approval of Class Action Settlement and Judgment, July 20, 2026.US District Court, ND California, Docket 680; a formal procedural outcome of a settlement agreement, not a general decision on the merits of all legal issues.Open the original source

[23] Kadrey v. Meta Platforms, Inc., Order on Cross-Motions for Partial Summary Judgment, June 25, 2025.US District Court, ND California, Docket 598; text of the decision (in addition to the WIPO review [13]).Open the original source

Editorial disclaimer: This text contains research proposals. The lack of actual access to the closed corpus, source models, or contracts of a specific developer precludes any conclusions regarding individual violations or the complete legality of their practices. The revision of October 23, 2026, includes a review of the legal chronology, selection of key official sources, arithmetic examples, and Word/PDF display. It does not certify the legality of specific closed training datasets. A separate, specialized legal and scientific review is required before official approval.