
SCCR/49/6
ORIGINAL: ENGLISH
DATE: OCTOBER 23, 2026

Standing Committee on Copyright and Related Rights

Forty-Ninth Session

Geneva, November 30 to December 4, 2026

SCOPING STUDY ON AI TRAINING AND COPYRIGHT

commissioned by the Secretariat

Scoping Study on AI Training and Copyright

DISCLAIMER:

This study was commissioned by the Secretariat of the World Intellectual Property Organization (WIPO). The views expressed in this document are those of the authors and are not intended to reflect the opinions of the Member States or of the Secretariat.

ACKNOWLEDGEMENT

DANIEL J. GERVAIS is the lead co-author of this study. He is the Milton R. Underwood Professor of Law (emeritus) at Vanderbilt Law School and a Senior Associate Professor at Vrije Universiteit Amsterdam and an elected member of the American Law Institute. His research can be accessed via www.collectivemanagement.net, www.tripsagreement.com and www.thehumancause.org.

SUNG JAI CHOI is Professor of Law at Sejong University. He holds an LL.M. from Columbia Law School. His research focuses on intellectual property and competition law. He is a co-author of Copyright Law (2024) and received the 2021 Korea Arbitration Grand Prize as a Next Generation Leader.

MELISSA OMINO is Director of the Centre for Intellectual Property and Information Technology Law (CIPIT) at Strathmore University in Nairobi. She holds a doctorate in intellectual property and trade law and specializes in intellectual property, artificial intelligence, data protection, and digital rights in the Global South.

RODRIGO KOPKE SALINAS is Professor at FGV São Paulo Law School (GVlaw) and holds an LL.M. from Columbia University. He is a member of the Special Council on Copyright of the Brazilian Bar Association and specializes in media and entertainment law.

WANG QIAN is Professor of Law and Vice Chairman of the Academic Committee at East China University of Political Science and Law. He is also Vice Chairman of the China Copyright Association and served as a member of the Expert Committee for the third revision of China's Copyright Law.

NOTE: Language Models were used to check the syntax, grammar, and contents of the study and to make editorial suggestions to harmonize this coauthored text.

TABLE OF CONTENTS

Introduction	4
Part I. The use of protected content in the development of generative AI models	5
Section A. Obtaining the data	5
1. Patterns of data consumption by system type	5
2. Web scraping	6
3. Retrieval-augmented generation	7
Section B. Using the data in training	8
1. Training the model	8
2. Optimization and reinforcement learning	10
Section C. After the training	10
1. Further uses of the training corpus	10
2. Retention of copies in the model	11
3. The input-output relationship	14
Part II. Authorization, Licensing, and Compensation	15
Section A. Current options for developers to obtain copyrightable subject matter for training	15
1. The five acquisition routes	15
2. Collaborative arrangements	17
Section B. Main policy elements	17
1. Rights affected by the training of generative AI models	17
(a) The reproduction right	17
(b) The right of adaptation and derivative works	18
(c) Communication to the public and making available	19
(d) Moral rights	19
(e) RMI and anti-circumvention	20
2. Voluntary (collective and individual) licensing	20
(a) Individual and collective licensing of AI uses	20
(b) The main collective licensing models (voluntary, extended, and compulsory)	24
3. Community and stewardship licensing of primary language data	28
4. Statutory TDM exceptions and general clauses	29
5. Remuneration systems	33
(a) Types of remuneration systems	33
(b) Modalities and basis of remuneration	34
Part III. Enforcement	35
Section A. Lessons from current case law	36
1. The model-as-copy question	36
2. The text and data mining exception	38

3. Jurisdiction over extraterritorial training	39
4. The substitution question	41
Section B. Enforcement against AI developers	44
1. Structural features of the enforcement landscape	44
2. Asymmetries between the parties	45
Section C. Challenges in the enforcement of rights	45
1. Transparency	45
2. Data attacks, filtering, and the proliferation of opt-out standards	47
3. The case law as a whole	50
Conclusion	51

INTRODUCTION

The present study has been prepared at the request of the Standing Committee on Copyright and Related Rights (SCCR). At its Forty-Seventh Session (Geneva, 1–5 December 2025), the Standing Committee requested the Secretariat to prepare a study on policy or regulatory approaches to the relationship between AI training and copyright, and related rules, and on applicable practices on authorization, enforcement and compensation for use. The request is recorded in the Summary by the Chair of that session (document SCCR/47/SUMMARY¹ of 5 December 2025, paragraph 21).

The study is descriptive and comparative in nature. It reviews the methods adopted in a number of jurisdictions and the practices that are developing and discusses the options that have been proposed without recommending any specific approach to the Member States or taking a position on the disputed issues that they have yet to resolve.

The study is about the training of generative AI models, i.e. the use of content for the building and improvement of such models and the questions of authorization, enforcement and compensation raised by such use when the content is protected by copyright or related rights. The larger intersection of artificial intelligence and copyright goes well beyond training. In particular it addresses the legal status of the outputs of a deployed model, both in terms of whether any particular output infringes the works on which the model was trained, and in terms of whether and on what conditions an output may itself be copyrightable. The questions relating to output raise other issues which the Standing Committee did not ask this study to address. They are only mentioned where it is necessary to give a full picture of the training question or to give necessary context.

The study is presented in three Parts. Part I describes, at a technical and descriptive level, the operations through which protected content is used in the development of a generative AI model and identifies the points at which those operations may implicate the exclusive rights recognized by national copyright legislation and international copyright treaties. Part II looks at the scope and extent of rights engaged and mechanisms by which they are used: voluntary licensing, individual or collective; and applicable exceptions and limitations. Part III deals with enforcement and explores the lessons that may be learned from the corpus of case law that has started to develop in a number of jurisdictions. It also addresses the cross-border nature of training and related enforcement issues, including transparency and standards for expressing opt-outs where applicable. The jurisdictions cited in this study are for illustration and are not exhaustive. They include Brazil, Canada, China, the European Union and many of its Member States, Kenya, the Republic of Korea, the United Kingdom and the United States.

¹ See https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=653196

PART I. THE USE OF PROTECTED CONTENT IN THE DEVELOPMENT OF GENERATIVE AI MODELS

A generative artificial intelligence (AI) model is developed in successive stages and third-party content is typically used in each of the stages. Some of the content may be subject to copyright and related rights; some of it may be public domain either because of the expiration of the term of protection or because it was never protected to begin with. The technical operations applied to such content are generally the same regardless of its legal status. The same processes, including crawling, copying, tokenization, and statistical analysis, which are explained below, may be applied equally to a recent novel and to an eighteenth-century text in the public domain. This Part describes these operations and indicates, at a descriptive level, the points at which they may implicate rights recognized by national copyright legislation and by international copyright and related rights treaties.

Three phases of training are presented in turn. Section A discusses data acquisition, i.e. how the AI model developer builds the corpus, often consisting in large part of protected subject matter, on which to train a model. Section B covers the actual use of that data in training, including the main training run and the different types of fine-tuning afterwards. Section C deals with the situation after training has taken place, including the treatment of the training corpus, and the question of whether the trained model itself contains copies of the works² on which it was trained.

SECTION A. OBTAINING THE DATA

The development of a large language model starts with the construction of a corpus, typically at very large scale and of mixed provenance. Developers obtain the data in several ways. Some data may be licensed, in which case the content may be made available together with the license to use it. In such a scenario, a developer can purchase access to a curated dataset compiled by a third party, or negotiate a direct licensing agreement with a publisher, a news organization or a collective management organization (CMO). If no license is granted for the use of the protected works, the use may be based on the assumption that an exception or limitation to the relevant exclusive right applies. A major training route, in terms of volume and the copyright issues it raises, is the automated retrieval of material from the open web (Internet). Training may also employ synthetic data, that is, data generated by an existing AI model rather than gathered from human creators, especially in later stages of development.

1. Patterns of data consumption by system type

A system's demand for data is not usually limited to the original training run: data are used for retraining and fine-tuning, for the text and data analysis done on a maintained search library, for retrieval-augmented generation (RAG)³, and for the generation of previews and snippets. These uses vary in the amount of material they require and in the ease with which their sources can be traced. The type of system tends to influence the patterns of consumption: media generation systems and chatbots based on large language models consume data mostly from training ingestion; customizable systems built by enterprises on top of a language (also called foundation) model add a domain-specific corpus and often a search library consulted at the point of use; and conversational search engines require access to online sources at the time of generation. Where a system falls on this spectrum

² To simplify the text, the term "works" should be understood to include objects of related rights where appropriate.

³ Retrieval-augmented generation, commonly abbreviated RAG, is discussed further in section 3 of this part.

influences the set of exclusive rights its use of data implicates and the corresponding structure of any license. This is developed in Part II, Section B.2.

2. Web scraping

Web scraping is a common approach employed by foundation-model developers to construct their training corpus. The terminology is not universally adopted, however. Some sources differentiate between “web crawling” (the automated movement through websites by a program that follows links from page to page) and “web scraping” (the selective pulling of identified material from a page that has already been retrieved). In practice the two activities may be combined: a developer deploys a crawler that follows links across the web and at each page identifies and extracts the textual, visual or other content that is to be retained. The output is often a copy of that content, generally stored in a structured manner on the developer’s own infrastructure.

Four features are relevant to the present discussion. First, the technique involves the making of copies. Copies may be retained on a server or in a cache or processed in memory before further transformation. These operations may implicate the reproduction right recognized in Article 9(1) of the Berne Convention, in Article 1(4) of the WIPO Copyright Treaty (WCT), and in corresponding provisions of national legislation. The agreed statement on Article 1(4) of the WCT confirms that the storage of a protected work in digital form is a reproduction and that the reproduction right and the exceptions thereto fully apply in the digital environment.⁴ Whether the temporary character of certain copies, or the operation of an exception such as that for incidental and transient reproductions, alters the analysis is a question taken up later in this study. Article 7 of the WIPO Performances and Phonograms Treaty (WPPT) and the same Article in the Beijing Treaty on Audiovisual Performances also contain a right concerning direct and indirect reproductions “in any manner or form.”

Second, retrieval may be non-selective. A crawler does not always distinguish between content that is in the public domain, content that is offered under an open license, content that is made available without any license statement at all, and content that is protected and is made available subject to terms of service or behind a technological protection measure. For example, some but not all crawlers respect the robots.txt instructions.⁵ Distinctions of this kind are sometimes drawn after the fact, through filtering or deduplication of the corpus, but the initial act of copying is not always selective.

Third, the visibility of the activity to rights holders is limited. A rights holder may not know that their work has been scraped, by which crawler, when, or for what subsequent use. The disclosure obligations that are beginning to appear in some national laws, in particular in the European Union under the Artificial Intelligence Act of 2024 (AI Act), require providers of general-purpose AI models to publish a “sufficiently detailed” summary of the content used for training.⁶ It is too soon to assess the practical effect of those obligations. The question of

⁴ WIPO Copyright Treaty (20 December 1996), Agreed Statement concerning Article 1(4): “The reproduction right, as set out in Article 9 of the Berne Convention, and the exceptions permitted thereunder, fully apply in the digital environment, in particular to the use of works in digital form. It is understood that the storage of a protected work in digital form in an electronic medium constitutes a reproduction within the meaning of Article 9 of the Berne Convention.”

⁵ The Robots Exclusion Protocol, standardized as IETF RFC 9309 (2022), is a convention by which a website operator places a plain-text file (robots.txt) at the root of a site to signal which automated crawlers may access which parts of the site. Compliance is voluntary; the protocol has no enforcement mechanism.

⁶ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 53(1)(d) (obligation, for providers of

how disclosure can be combined with respect for the legitimate commercial interests of developers is itself the subject of ongoing discussion.⁷

Fourth, the making and use of copies of protected works for training may involve the removal or alteration of rights management information and the circumvention of effective technological measures that are used by right holders in connection with the exercise of their rights under the WCT, WPPT and Beijing Treaty.

3. Retrieval-augmented generation

A separate but related technique is retrieval-augmented generation (RAG).⁸ It is a method by which an AI system answers a query by first consulting a separate, pre-assembled collection of documents rather than relying only on the material on which it was trained. A useful comparison is the difference between a person answering a question from memory and the same person pausing to look something up in a reference shelf kept beside the desk. When a query arrives, the system searches that collection, retrieves the passages most relevant to the question, places them alongside the query, and uses them to compose its answer. The collection is built in advance by making and storing copies of the documents, and the relevant passages are often copied again into the system at the moment the answer is generated, points that matter for the copyright analysis set out below.

In a RAG system, an external corpus is maintained, typically indexed in a database after each document or document fragment has been converted what is known as an embedding⁹. When the system receives a user query, it is embedded and used to retrieve the most relevant fragments from the corpus. The system then assembles the prompt that is fed to the language model, combining the user's query with the retrieved fragments, which the model uses to generate output based on its training and the retrieved material.

For the present purposes, RAG is sometimes described as modifying the copyright questions that arise from language-model training. The reasoning is that the model itself does not learn from the corpus because the corpus is held separately and consulted at inference time. The construction of the indexed corpus requires the prior processing and, in a conventional RAG architecture, storage of the source documents or their constituent fragments.¹⁰ The conversion of each document into an embedding produces a numerical

general-purpose AI models, to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office").

⁷ Occasionally, however, the provenance of scraped material surfaces in the behavior of the resulting model: early versions of a prominent image-generation model reproduced a recognizable form of a stock-photography agency's watermark, an indication that a large volume of that agency's images had been ingested without authorization. The phenomenon was widely reported and featured in the litigation discussed in Part III, Section A.1 (Getty Images (US) Inc. v. Stability AI Ltd.). It illustrates that, although the scraping of a particular work is generally invisible to the rights holder, the trained model may itself reveal, through its outputs, that protected material was used in training.

⁸ See note 2 above.

⁹ Word embeddings represent each token as a dense high-dimensional vector of real values. Embeddings are described as "contextual" when the representation assigned to a given token depends on the tokens that surround it. See Heart of the Matter, above note 10, at Part II.C.

¹⁰ In more technical terms, a distinction should be drawn between the passage embeddings used for vector search and the source text from which they are generated. An embedding is a dense numerical representation produced by an encoder; it is not an exact copy of the text and is not ordinarily reversible by simply decoding the vector. Specialized "embedding inversion" techniques can nevertheless reconstruct some or, under particular experimental conditions, all of relatively short source passages from their embeddings. In a conventional RAG system, the source documents or fragments are also retained as text, either in the vector store itself or in a separate document store linked to the vectors. The embeddings are used to identify relevant passages; the

representation derived from the document, although, unlike the stored source copy, the embedding is generally not reversible to the original text. At runtime (inference), the relevant fragments are copied into the model's context window and may be reproduced, in whole or in substantial part, in the output. These steps implicate the reproduction right; and could implicate the adaptation right (Berne Convention, Article 12) depending on the applicable law and the characteristics of the representation. The possible application of those rights is discussed below.

SECTION B. USING THE DATA IN TRAINING

Training a generative AI model is a sequence of operations commonly described as having four stages: pretraining, supervised fine-tuning, reward modeling, and reinforcement learning.¹¹ The first of these is generally the most demanding, both in computational cost and in the volume of data consumed; the later three stages, taken together, are often described as "fine-tuning" or, sometimes, as "post-training." A separate but related activity, model evaluation and optimization, accompanies the training and continues into the deployment of the model.

For copyright purposes, the principal observation is that each of these stages usually involves the storage and processing of all or part of the underlying corpus. The training pipeline is not one in which the data are consulted briefly and then discarded; the data are stored, loaded, transformed, and used to update the model. Where the data include protected works, those operations involve copies or transformations relevant to the reproduction analysis developed in Part II.

1. Training the model

The principal training run, generally termed pretraining, is the stage in which the model acquires the bulk of its knowledge of language (in the case of a language model) or of images, audio, and video (in the case of a multimodal model). Pretraining proceeds in several technical steps, each of which is performed on a stored copy of the corpus. The textual content is first subjected to tokenization, which is the conversion of words and other sequences into a stream of numerical tokens drawn from a fixed vocabulary.¹² Modern tokenizers generally operate at the level of subwords rather than whole words.¹³ The step matters for the present analysis because, in the case of the byte-level tokenizers used by the

associated text, rather than the embedding vector, is then retrieved and supplied to the language model as part of its input context. See John X. Morris et al., "Text Embeddings Reveal (Almost) As Much As Text", Proc. 2023 Conf. on Empirical Methods in Natural Language Processing (EMNLP) (2023). For an industry description of the retrieval pipeline, see:

<https://infohub.delltechnologies.com/en-us/p/demystifying-on-device-intelligent-search-using-rag-architecture/>

¹¹ The four-stage description used here is drawn from the technical literature and is discussed at greater length in *The Heart of the Matter: Copyright, AI Training, and LLMs* 71 J Cop Soc'y of the USA 482 (2024) (hereinafter *Heart of the Matter*), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4963711. A similar four-stage pipeline (data collection, data preprocessing, model training, and model evaluation and optimization) is presented in the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission, *Guide on Applicability of the Fair Use Doctrine to Training of Generative AI Models* (February 2026) (hereinafter *Korea Fair Use Guide*), at 10.

¹² On tokenization, see *Heart of the Matter*, above note 10, at Part II.B. The *Korea Fair Use Guide* refers to tokenization as one of the technical procedures of data preprocessing alongside cleaning and normalization. See *Korea Fair Use Guide*, above note 10, at 10.

¹³ The vocabulary is built by starting from individual characters and progressively merging the most frequently occurring adjacent pairs into larger units, so that common words come to be represented as single tokens while rarer words are split into several. Byte-pair encoding is used by OpenAI for its Generative Pretrained Transformer (GPT) models and by a number of other major developers. See *Heart of the Matter*, above note 10, at Part II.B.

principal contemporary models, it is lossless and reversible: text encoded by such a tokenizer and then decoded yields the original text, so the tokenized corpus is not a summary or an abstraction of the works but a re-encoding of them in another form.

The tokens are then represented as embeddings.¹⁴ The approach that has dominated the field since 2017, the Transformer (the “T” in the acronym GPT, which stands for Generative Pre-trained Transformer), processes those representations through a mechanism known as attention.¹⁵ The characterization of the embedding, and of the trained model itself, which involve a greater degree of transformation, is taken up in Section C.2 below.

Once the model has been initialized, it is trained by being repeatedly exposed to a sequence of tokens drawn from the corpus. The model predicts each token in turn based on the tokens that precede it and is adjusted, in many small increments, so that its predictions come closer to the observed text; the corpus is thus not consulted once but used, token by token, throughout the training run. After processing an enormous volume of tokens drawn from the corpus, the model has internalized a massive probability distribution over a vast space of possible token sequences. That distribution, encoded in the model’s parameters, is what allows the model to generate coherent text in response to a prompt.

The relationship between the training data and the trained model is the subject of substantial debate. One position is that because the trained model contains only parameters and not the original training data, no copies of protected works remain after training is complete.¹⁶ The parameters describe a probability distribution over an enormously large vector space. While this distribution allows the model to generalize and generate entirely novel sequences, it can assign disproportionately high, concentrated probabilities to specific, exact sequences that were repeatedly encountered during training. Because the distribution is heavily weighted toward these finite training paths, the model can, in many cases, be induced to reproduce them verbatim or near-verbatim. The phenomenon is widely documented in the technical literature and is discussed further in Section C below. There is, as explained below, a difference between this phenomenon (often referred to as “memorization”) and the issue of whether a “reproduction” exists within the model of all works used for training.

For models that generate images, audio, or video, the most prevalent models are known as “diffusion”. They learn by progressively adding noise to training examples and then learning to reverse that process so as to produce new examples from noise¹⁷. This is different from language (text) models, but in both cases the training process produces a sophisticated representation of the input data.

¹⁴ More specifically, they are mapped into a high-dimensional vector space, in which each is represented as an embedding.

¹⁵ The attention mechanism allows the model, in computing the representation of a given token, to weigh the relevance of every other token in the sequence, which is what enables it to capture dependencies across long passages of text. Ashish Vaswani et al., “Attention Is All You Need”, 31 Conf. Neural Inf. Processing Syst. (NeurIPS) (2017). The Transformer architecture underlies the principal contemporary large language models.

¹⁶ See Heart of the Matter, above note 10, at Part II.C.

¹⁷ On the origins of diffusion models, see Jascha Sohl-Dickstein and others, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”, Proc. 32nd Int’l Conf. on Machine Learning (ICML) (2015); Jonathan Ho, Ajay Jain and Pieter Abbeel, “Denoising Diffusion Probabilistic Models”, 34 Conf. Neural Inf. Processing Syst. (NeurIPS) (2020); and, on the latent diffusion architecture used in Stable Diffusion and at issue in the Getty litigation discussed in Part III, Robin Rombach and others, “High-Resolution Image Synthesis with Latent Diffusion Models”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). For an accessible introduction, see also Yang Song and Stefano Ermon, “Generative Modeling by Estimating Gradients of the Data Distribution”, 33 Conf. Neural Inf. Processing Syst. (NeurIPS) (2019) available at <http://papers.neurips.cc/paper/9361-generative-modeling-by-estimating-gradients-of-the-data-distribution.pdf>. The general point in the present discussion does not turn on the mathematical details of any particular class of model.

2. Optimization and reinforcement learning

The stages following pretraining are responsible for adapting the pretrained model to particular tasks, to particular use cases, and to particular standards of behavior. The terminology varies across developers, but the sequence includes supervised fine-tuning, reward modeling, and reinforcement learning.¹⁸

Supervised fine-tuning involves further training on a smaller, curated dataset that consists of input-output pairs of the kind the model is expected to produce. The dataset is much smaller than the pretraining corpus, but it is selected for quality and is often human-crafted or human-edited. Each fine-tuning example, like each pretraining example, is loaded and used to update the model's parameters.

Reward modeling involves the construction of a separate "reward" model that scores candidate outputs from the principal model according to their desirability. The reward model is itself trained on human-generated comparisons of candidate outputs.

Reinforcement learning from human feedback (RLHF), a widely used technique for aligning generative AI models with human preferences, uses the reward model to guide further updates to the principal model.

Each of these stages requires the storage and processing of training material. The volumes are smaller than those required for pretraining, but the training material may itself include protected works, particularly where the comparisons supplied by human annotators are based on or include such works. Focusing on the quality rather than the quantity of the material does not alter the analysis from a copyright perspective.

Model evaluation and optimization, which run alongside the training stages and continue after deployment, involve the use of evaluation datasets to assess the model's performance and to identify areas in which further training is needed. Evaluation datasets may themselves include protected works, particularly where benchmarks have been constructed from publicly available test corpora.

SECTION C. AFTER THE TRAINING

Three distinct questions arise once the training is complete. The first is whether the corpus is retained, by whom, and for what subsequent purposes. The second is whether and in what sense the model contains copies of the works on which it was trained. The third is whether outputs reproduce, adapt, or otherwise engage with the works that were used in training.

1. Further uses of the training corpus

The training corpus remains relevant after the training run is complete. It may be retained by the developer for several purposes: to evaluate future versions of the model, to support further fine-tuning on particular subsets, to enable ongoing research, or to be made available to third parties under license or otherwise. It may be combined with newly assembled material to form an enlarged corpus for the next generation of models. It may be sold or shared (e.g., with affiliated organizations). In some cases, derivatives of the corpus, including filtered or deduplicated subsets, are released into the public domain under permissive licenses.

Each of those subsequent uses may involve acts of reproduction, distribution, and making available within the meaning of the relevant national and international rules. In other words, a corpus that has fulfilled its training purpose may continue to be used for years afterward,

¹⁸ See Heart of the Matter, above note 10, at Part II.D. The terminology of "pretraining," "supervised fine-tuning," "reward modeling," and "reinforcement learning" is in widespread use among AI developers.

and the rules governing its subsequent use align with those that governed its initial assembly.

2. Retention of copies in the model

Whether the trained model itself contains “copies” of the works on which it was trained is one of the most technically demanding questions discussed in this Part and one that differentiates generative AI from earlier technologies. The positions taken by AI developers, on the one hand, and by rights holders, on the other, have until recently been sharply opposed.¹⁹ AI developers have generally argued that the trained model contains only parameters, not the works themselves, and that no copies are retained for purpose of copyright law. Rights holders have argued that the model is, in effect, a copy of the entire corpus. The empirical evidence accumulated in the past two years suggests that both positions tend to oversimplify the relationship between memorization and the model. Moreover, whatever the resolution of that dispute, it does not affect the anterior acts: the assembly of the corpus and its use in training involve reproductions whose legal characterization is independent of whether the completed model itself contains copies.

Two lines of technical evidence can be used in the analysis. Memorization is a property of the trained model which can recover specific training sequences in the model’s parameters with sufficient fidelity. Because the internal structure of a model cannot easily be inspected directly, memorization is studied through extraction, that is, the demonstrated ability of the model, under suitable prompting, to reproduce sequences from its training data, verbatim or near-verbatim. Extraction is then used as evidence of memorization. The failure of a given extraction procedure does not prove that a work has not been memorized. A systematic study, by Cooper and others applied a probabilistic extraction technique to 200 books across fourteen open-weight language models and found that the extent of memorization varied significantly both by model and by book.²⁰ Most of the largest models did not memorize most of the books, in whole or in part. Where memorization occurred, however, it could be substantial: the study reported that one widely deployed model had memorized the first Harry Potter novel and George Orwell’s 1984 almost entirely; in the case of the former, the study demonstrated that essentially the entire book could be generated near-verbatim.

Those empirical demonstrations of memorization are not, however, uncontested. AI developers have responded by arguing that the prompting techniques required to induce verbatim reproductions are “adversarial,” are not representative of how the systems are used in practice, and may themselves constitute a manipulation or misuse of the product that falls

¹⁹ The divergence extends beyond litigation positions to design priorities. A multi-stage Delphi study among AI ethicists, legal experts, music AI practitioners, and artists, conducted to validate a responsible-design framework for generative music systems, found a “protectionist” alignment between artists and legal experts on questions of data sovereignty, contrasted with the feasibility focus of practitioners, and concluded that such frameworks operate best as a translation layer between divergent stakeholder priorities, including the trade-off between model transparency and intellectual property interests, rather than as instruments of forced consensus. J. de Berardinis and others, “RAIM: Operationalising Trustworthy AI in Music Generation”, IEEE Transactions on Technology and Society, Special Issue on Ethical Innovation with/in Music Technology (forthcoming, 2026), framework website <https://casbahlab.github.io/raim-framework/> accessed 25 August 2026.

²⁰ Feder Cooper, Aaron Gokaslan, Ahmed Ahmed, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho and Percy Liang, “Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models”, arXiv:2505.12546 (2025), version of 30 November 2025 (also published as Stanford Public Law Working Paper / WVU College of Law Research Paper No. 2025-005). The authors’ framing of the question (at p. 1) bears quotation in full: “Plaintiffs and defendants in copyright lawsuits over generative AI often make sweeping, opposing claims about the extent to which large language models (LLMs) have memorized plaintiffs’ protected expression. Drawing on both machine learning and copyright law, we show that these polarized positions dramatically oversimplify the relationship between memorization and copyright.” On the proposition that memorized content must be encoded in the model parameters because “there is nowhere else it could be,” see also A. Feder Cooper and James Grimmelman (cited therein at p. 25).

outside ordinary use (and, in some cases, Terms of Service). In the litigation brought by The New York Times against OpenAI, the defendant moved to dismiss the direct copyright infringement claims in part on the basis that the plaintiff had used what the defendant characterized as elaborate and manipulated prompts to manufacture the regurgitation it complained of.²¹ The District Court for the Southern District of New York declined to dismiss the claims on that ground in April 2025, and the dispute has since shifted, in the discovery phase, from whether “regurgitation” can be manufactured to how often it occurs organically in practice, culminating in a ruling compelling the production of a representative sample of 20 million real-world user interaction logs.²² The litigation, and the significance of the order, are discussed further in Part III, Section A.4, of this study.

In summary, the empirical findings complicate both polarized positions: the trained model is not, as a general matter, a “copy” of its training corpus, but neither is it “free” of the corpus. The exact scope will vary by model.

The second line of evidence is the broader literature on data extraction attacks, which documents circumstances in which individual training examples can be recovered from a trained model through targeted prompting.²³ The recovered examples include text passages of substantial length, identifiable images, and personally identifiable information. This work has been cited as evidence that extractability can be a structural property of large generative models rather than solely a result of user manipulation. Some AI companies argue that such prompting constitutes an unfair manipulation or a violation of terms of service.²⁴ Assessments of this kind have, however, become more complex over the past year: deployed systems increasingly interpose filtering and other guardrails between the model and the user to reduce extractability of memorized content; recent studies have had to resort to adversarial techniques for some systems, while others yielded protected content without them.²⁵

An analogy may be useful to understand the situation. It is an analogy with computer programs²⁶. When source code for a program (often written in a language resembling a

²¹ The New York Times Company v. Microsoft Corp., No. 1:23-cv-11195, Memorandum in Support of Defendants’ Motion to Dismiss (S.D.N.Y., 26 February 2024); Opinion and Order (S.D.N.Y., 4 April 2025) (Stein J.) (denying in relevant part the motions to dismiss the direct and contributory infringement claims). For OpenAI’s own public framing of the argument, see OpenAI, “Reporting the Facts About the New York Times’ Lawsuit,” available at openai.com/new-york-times.

²² In re OpenAI, Inc. Copyright Infringement Litigation, No. 1:25-md-03143-SHS-OTW (S.D.N.Y.), Opinion and Order (Stein J., 5 January 2026), affirming Opinion and Order (Wang M.J., 2 December 2025) (denying OpenAI’s motion for reconsideration) and the underlying discovery order of November 2025. The court ordered the production of a randomised, de-identified sample of 20 million ChatGPT user interaction logs, drawn from the preserved population, on the basis that broad consumer-facing output evidence was relevant to the fair use analysis, in particular to the question whether the model’s outputs function as a market substitute for the plaintiffs’ works.

²³ Earlier work in the same line includes Antonia Karamolegkou et al., “Copyright Violations and Large Language Models”, arXiv:2310.13771 (2023) (demonstrating, across a range of language models, that individual training data examples can be recovered through data extraction attacks).

²⁴ See Nicholas Carlini et al., Extracting Training Data from Large Language Models, 30th USENIX Security Symposium (2021) 2633; Milad Nasr et al., Scalable Extraction of Training Data from Aligned, Production Language Models, ICLR 2025 (extending Nasr et al., arXiv:2311.17035 (2023)).

²⁵ Ahmed Ahmed and others, “Extracting Books from Production Language Models”, arXiv:2601.02671 (2026) (finding that adversarial prompting was required to elicit memorized book text from some deployed systems but not from others).

²⁶ See Heart of the Matter, above note 10, at Part II.C (drawing the analogy between the training of a language model and the compilation of a computer program into executable form, and noting that, in both cases, “the original copyrighted work” is “translated into another form that is not directly understandable by humans” but in which the content of the original “is [not] absent from that machine-understandable version”).

natural language) is compiled into executable form, the result is different from the source code and is not (easily) readable by a human. The compilation is not, however, free of the original: the executable retains the information of the source program, and, by reverse engineering, the original can be reconstructed to a substantial extent. To that extent, the executable code remains similar to the source program. The steps that produce a trained model from a training corpus are, if one follows this analogy, arguably of a comparable kind. In both cases, an intelligible source (a program in source code; a corpus of works) is transformed into a functional numerical object that no longer resembles it but continues to carry information derived from it. The analogy has limits, however: compilation is a deterministic process that carries the entire source program into the executable, whereas the encoding of training data in a model is selective and in part probabilistic, so that the analogy holds, at most, for the subset of works shown to be memorized; recent scholarship argues that this difference between deterministic and merely probabilistic recoverability is precisely where the legal characterization of the model is likely to turn.²⁷

The technical evidence reviewed here indicates that a trained model retains information derived from works in its training data even though the transformation is substantial and access to any particular work may be partial and probabilistic. How this technical observation translates into the legal characterization of the model and into the question of whether the model itself is a copy or an adaptation/derivative of the works on which it was trained as a matter of copyright law is the subject of ongoing analysis in the courts and in the scholarly literature, as this study explains later on.

Two observations may be drawn at this juncture. First, memorization is being used as an argument both at the training stage, where it is used as evidence that training the model involved the encoding of copies, and at the output stage, as evidence that the deployed model can generate protected works in a form that substitutes for the originals. Second, the demonstrations on which the plaintiffs rely concern particular works in particular models. They tend to confirm that memorization can be substantial for specific, individual, frequently encountered works but do not establish that it is a uniform property of all models or that a model's parameters reduce in general to stored copies. The legal characterization of these questions is taken up in Part III.²⁸

Finally, it is also worth noting that a model may change once training is complete through fine-tuning, alignment training, or attempts at machine “unlearning” directed at specific content. A model that ceases to reproduce a given work after such an intervention has not thereby been shown never to have encoded it. The technical literature indicates that reliably

²⁷ M.A. Lemley and A.F. Cooper, “Probabilistic ‘Copies’ in Generative AI Models”, Berkeley Technology Law Journal (2026), available at <https://ssrn.com/abstract=7067878> (arguing, from within the empirical extraction literature discussed in this Part, that a model should be regarded as containing a “copy” of a particular work only where the work can be extracted in a straightforward and, in the strongest cases, deterministic manner; that where extraction succeeds only probabilistically, after many attempts, what is stored is not a copy of the work but a set of “ingredients”; and that the question is accordingly evidentiary, to be answered work by work and model by model).

²⁸ On the consolidated proceedings, see *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143-SHS-OTW (S.D.N.Y.), consolidating the actions brought by The New York Times and by groups of authors against OpenAI and Microsoft. The characterization of the trained model as a storage and retrieval mechanism, and of the parameters as a compressed form of the ingested works, reflects the plaintiffs’ pleadings; the contrary characterization of the parameters as statistical generalizations encoding patterns of language reflects the defendants’ position. On the empirical literature cautioning against both characterizations, see the Cooper and others study cited above in this Section.

removing specific works from a trained model remains difficult, and that such interventions can degrade the model in unexpected ways.²⁹

3. The input-output relationship

Once the model is deployed, it operates by receiving a user input, or prompt, and producing an output. The output may be text, code, an image, audio, video, or another form of content. The relationship between the input and the output is in some respects analogous to the relationship between a query and a response in a search engine, but the analogy is partial. The model is not consulting an indexed corpus and returning passages from it; it is generating new content by sampling the probability distribution it has internalized during training. In some cases, what it generates is closely related to specific works in its training data; in other cases, the output is a novel composition that does not closely resemble any single training example.

Three observations are particularly relevant. First, the output may be a verbatim or near-verbatim reproduction of a work in the training data, in which case the output constitutes a reproduction of the original work.³⁰ This has been demonstrated in studies of language models reproducing passages from copyrighted books and of image-generation models reproducing protected photographs and artworks³¹. Whether such reproduction is held to infringe will depend on a number of further questions, including the applicability of exceptions and limitations, the role of any output filters deployed by the developer, and the attribution of responsibility between the developer of the model and the user who issued the prompt. Those questions are mentioned briefly later in this study³².

Second, the output may be substantially similar to a work in the training data without being a verbatim reproduction. This may constitute infringement under, for example, the substantial similarity doctrine, a familiar standard in a number of national copyright laws.

Third, the output may be entirely new, in the sense that no particular work in the training data is identifiable as its source. A generative model is capable of producing content that does not match any single training example. Indeed, in most cases that is precisely what it does. In such cases, the relationship between the output and any specific protected work is attenuated. Whether the output remains “based on” the training data in any sense that copyright law would recognize is a question that has been put to the courts in several jurisdictions but on which a consensus has not yet emerged.

Part I therefore identifies uses of protected works, and questions concerning their retention or reproduction, at multiple stages of the development and operation of a generative AI model: in the corpus that is assembled at the outset, in the storage and processing operations applied to the corpus during training, in the trained model itself, and in the outputs produced after deployment. At each stage, the principal exclusive rights recognized by national copyright legislation and by the international treaties may be involved. The question

²⁹ A. Feder Cooper and others, “Machine Unlearning Doesn’t Do What You Think: Lessons for Generative AI Policy, Research, and Practice”, 39 Conf. Neural Inf. Processing Syst. (NeurIPS) (2025).

³⁰ Where an LLM output is similar to a preexisting protected work that the model did not use for training and to which it did not otherwise have access, it is much less clear that a “copy” has been made.

³¹ Verbatim reproduction by language models has been documented in academic studies and in litigation filings; reproduction of recognizable photographs and artworks by image-generation models has been documented in connection with diffusion models. The volume of documented examples does not bear directly on the legal characterization of any particular output, which depends on the applicable national law and on the facts of the case.

³² On the relationship between the input prompt, the model, and the output, and on the open question of how responsibility is to be allocated between the developer and the user, see Part III below.

whether each reproduction may amount to a prima facie infringement (subject to an exception or limitation), and the question of what mechanisms exist or could be developed to license the relevant uses, are taken up in the Parts that follow.

PART II. AUTHORIZATION, LICENSING, AND COMPENSATION

Part I of this study identified, at the technical level, the points at which the principal exclusive rights recognized by national copyright legislation and by the international copyright treaties may be implicated in the development and operation of a generative AI model. The present Part turns from those rights to the mechanisms by which they may be exercised, waived, or supplanted in the AI context. The question whether the relevant uses are lawful without authorization and, where they are not, the question of how developers can place their operations on a lawful footing and how rights holders are to be compensated through that process, lead to a single body of doctrine: the law of copyright licensing, individual and collective, voluntary and statutory, together with the exceptions and limitations that condition the need for a license in the first place.

The Part is organized as follows. Section A discusses the main channels through which AI developers currently acquire the copyrightable subject matter on which they train their models and the legal position of each channel. Section B addresses the principal policy issues currently shaping licensing practices: the rights affected by training, voluntary licensing, collective licensing in its three main forms, the status of the EU text and data mining exception and the scholarly discussion on its relevance to AI training, and remuneration schemes. Countries are mentioned as examples of approaches adopted in different parts of the world; they are not the object of the analysis and the survey is not exhaustive.

SECTION A. CURRENT OPTIONS FOR DEVELOPERS TO OBTAIN COPYRIGHTABLE SUBJECT MATTER FOR TRAINING

1. The five acquisition routes

In practice, AI developers acquire their training data in one of five ways, or some combination thereof. One is the use of open or curated data sets created by third parties and offered, sometimes free of charge, sometimes for a fee. A prominent example is Common Crawl, a continuously updated repository of webpages, crawled and made available by a non-profit foundation, and which underpins, directly or indirectly, the textual training corpora of most major language models. LAION-5B and its predecessors are the equivalent for image-generation models.³³ Such datasets are themselves the result of large-scale automated retrieval operations and may carry forward copyright issues associated with the underlying retrieval. The legal position will depend on the basis on which the source material was obtained and on the law applicable to subsequent uses of the dataset. A public-sector counterpart to these curated datasets has begun to appear: in the Republic of Korea, the public data license system (Gonggong Nuri) has been updated to include a license category

³³ Common Crawl is described at <https://commoncrawl.org>. The LAION-5B dataset (used in the construction of Stable Diffusion and other image-generation models) and its predecessor LAION-400M are publicly available image-text datasets compiled by the Large-scale Artificial Intelligence Open Network. The dataset itself does not contain the underlying images but pairs of URLs and textual descriptions; the images are retrieved at training time from their original locations. On the legal characterization of dataset compilation operations as text and data mining, see *Kneschke v. LAION e.V.*, LG Hamburg, 27 September 2024 (310 O 227/23), and on appeal, *Kneschke v. LAION e.V.*, OLG Hamburg, 10 December 2025 (5 U 104/24).

under which public datasets carry an express indication that their use for AI model training is authorized.³⁴

Another source is the so-called “shadow libraries”, large unauthorized repositories of books, articles, music and other works, which a number of widely used models are reported or alleged to have used, either directly or via datasets which include them. The United States piracy findings discussed in Part III, Section A, below, exemplify the legal exposure that reliance on such a repository engenders.

For the sake of comparison, some courts have distinguished between the illegal acquisition of works and the separate question of whether the use of the works in training is covered by an exception or limitation. In one of the US rulings discussed below, the downloading and retention of pirated copies was held not to be fair use, whereas the training of a model on lawfully-acquired copies of the same works was held to be transformative and fair. The legality of the acquisition and the legality of the training are, in this sense, analytically independent. The distinction matters because training data may come from lawful and unlawful sources within the same model, creating complexity in licensing and enforcement. A developer may therefore need both lawful acquisition of the corpus and a lawful basis for the training itself, subject to applicable national law(s).

The third approach is individual and collective licensing. A developer can purchase access to a curated corpus or negotiate for it from a publisher, a news organization, a broadcasting body, a picture library, an academic database operator or any of the institutional rights holders. Licensing has become a familiar practice in the last three years, with major AI developers signing deals with major news organizations and image libraries in Europe, the Americas and Asia.³⁵ The structural feature of licensing is that the developer obtains an authoritative chain of title to the licensed corpus, at least as to the licensor’s own rights. The limitations of that chain are discussed in Section B.2 below. In some cases, the contract includes delivery of the content.

The fourth, and a major route in terms of volume, is the direct collection of material from the open web via web crawling and scraping performed by the developer itself, not a third party. The technical procedures are described in Section A.2 of Part I of this study. The magnitude of automated retrieval involved is considerable: by some estimates, automated agents now account for a majority of global web traffic.³⁶ Crawling is often conducted without the prior consent of the rights holders of the retrieved material, on one of three legal theories: that the material is in the public domain or has been validly licensed to the public on permissive terms; that the operation is covered by an exception or limitation in the applicable copyright law (in the European Union, the text and data mining exception in Article 4 CDSM; in the United States and, under Article 35-5 of the Copyright Act, in the Republic of Korea, a general fair use doctrine; in Singapore and Japan, specific statutory provisions); or that the operation does not engage the reproduction right at all because the copies made are temporary or transient, or that, in some jurisdictions, such copies fall within an express

³⁴ The reference is to the Gonggong Nuri (Korea Open Government License) system, under which public institutions make public data available for reuse under standardized license categories; the updated framework adds a category expressly permitting use for AI model training.

³⁵ On bilateral licensing deals between AI developers and publishers, see (among many) the agreements concluded by OpenAI with the Associated Press, News Corp, Le Monde, Prisa Media, Axel Springer, Vox Media, The Atlantic, and Hearst; by Google with Reddit and the Associated Press; and by Anthropic with various publishers. The deals are disclosed at varying levels of detail and are commercially confidential as to their financial terms.

³⁶ See Imperva, 2025 Bad Bot Report (twelfth edition, 2025), reporting that automated traffic accounted for 51 per cent of all internet traffic in 2024, surpassing human traffic for the first time in a decade, with so-called “bad bots” accounting for 37 per cent.

exception for temporary acts of reproduction (as under Article 5(1) of the InfoSoc Directive in the European Union). These theories are not mutually exclusive and a developer might use different theories for different parts of the corpus.

The fifth is synthetic data: data generated by an existing AI model and used to train or fine-tune another. Synthetic data is a growing tool, most often used in the later stages of model development, in particular supervised fine-tuning and reward modeling.³⁷ From a copyright perspective, synthetic data poses a different question. If the model that generated it was trained on protected works, do the synthetic outputs also reproduce or adapt protected expression? The question has not yet been put to a court in a way that would provide a binding answer.

2. Collaborative arrangements

Finally, in addition to these five broad avenues, in some jurisdictions we observe forms of organized collaboration between AI developers and institutional or state-owned rights owners most relevant to particular sectoral applications. China provides one example. The development of vertical-domain large models has seen collaboration between a developer of general-purpose AI models and Hong Kong radio and television stations (for Cantonese voice data), between developers and online publishers with large repertoires of contracted writers (for digital novels), between developers and large picture repositories with billions of images held by transfer or exclusive license (for visual training data), and between developers and academic database operators jointly building specialized models. The pattern is said to be widespread, although the agreements themselves are often not titled as copyright licenses and their precise scope is not always disclosed. As discussed in Section B.2 below, the structural limit of this approach is that the publisher, broadcaster or repository typically holds only particular rights granted by the underlying author, and it is itself uncertain whether the rights so held cover AI training.

SECTION B. MAIN POLICY ELEMENTS

1. Rights affected by the training of generative AI models

(a) The reproduction right

The technical operations described in Part I involve, *prima facie*, a number of the exclusive rights recognized by national copyright legislation and by the international copyright treaties. The reproduction right is central to the analysis.³⁸ The construction of a training corpus involves the making of copies; the storage and processing of that corpus during the various stages of training involves further copies, some persistent and some transient; certain aspects of the trained model, on the technical analysis set out in Part I, Section C.2, may themselves be analyzed as copies of the works on which the model was trained. Each of those operations falls, *prima facie*, within the reproduction right as recognized in Article 9(1) of the Berne Convention, in Article 1(4) of the WIPO Copyright Treaty, Article 7 of the WPPT and of the Beijing Treaty, and in the corresponding provisions of national legislation.

The treatment of temporary and transient copies requires separate consideration, because two distinct legal techniques lead to similar practical results: in some jurisdictions such

³⁷ Synthetic data is data generated by an existing AI model and used in the training or fine-tuning of another model. It raises distinct copyright questions where the generating model was itself trained on protected works. On its growing use across the later stages of model development, see R. Liu and others, "Best Practices and Lessons Learned on Synthetic Data", arXiv:2404.07503 (2024), presented at the First Conference on Language Modeling.

³⁸ See Part I, Sections A.1 and A.2.

copies are not treated as falling within the scope of the reproduction right, while in others they are accommodated through an express exception. China provides the principal example of the first technique: temporary reproduction has consistently been regarded as falling outside the scope of the right. In Chinese judicial practice, the temporary reproductions so excluded are the transient copies made in random-access memory or in a cache that are strictly necessary for a process of data handling; there is no settled position on precisely how long a copy may subsist and still be regarded as temporary. Article 10(1)(5) of the Copyright Law defines reproduction as the right to produce one or more copies of a work by means such as printing, photocopying, engraving, recording, videotaping, duplicating, photographing, or digitizing,³⁹ and the Chinese delegation, among others, voted against the agreed statement at the Diplomatic Conference that would have brought temporary digital storage clearly within the scope of the right.⁴⁰ The Regulation on the Protection of the Right of Communication to the Public on Information Networks (2006) did not include a provision recognizing temporary reproduction as an act of reproduction, and the Chinese courts have applied this understanding in cases involving transcoding and similar transient digital operations. From this perspective, the act of training itself, to the extent that it involves only transient access to works in memory or cache that disappears upon completion of the training run, may not engage the reproduction right under Chinese law, although the construction of the underlying database of works certainly could. In other jurisdictions, the same practical outcome is reached through an express exception: in the European Union, Article 5(1) of the InfoSoc Directive accommodates transient and incidental copies through a narrowly drawn exception rather than by excluding them from the right *ab initio*. The copies concerned are in substance the same in the two systems; the difference lies in the legal technique employed.⁴¹ In most other jurisdictions surveyed here, the act of training is generally regarded as engaging the reproduction right, and the analytical question turns on the availability of an exception or limitation.

(b) The right of adaptation and derivative works

In addition to the right of reproduction, three other rights are or may be involved. The right of adaptation, which is recognized in Article 12 of the Berne Convention and in the corresponding national provisions, is invoked in the output phase, as the generation of output is considered as producing a transformation or alteration of the underlying works. For example, Brazilian law expressly lists the right of adaptation among the forms of use that require prior and express authorization,⁴² and the Secretariat of Copyright and Intellectual Property of the Ministry of Culture has taken the position that the generation output engages

³⁹ Copyright Law of the People's Republic of China, Article 10(1)(5).

⁴⁰ See WIPO Doc. CRNR/DC/101, Summary Minutes, Plenary, paras 497-502. The Agreed Statement concerning Article 1(4) WCT was nevertheless adopted by the Diplomatic Conference; China voted against it.

⁴¹ Article 5(1) of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJ L 167/10 [often referred to as the "InfoSoc Directive"]. The exception covers temporary acts of reproduction that are transient or incidental, an integral and essential part of a technological process, and whose sole purpose is to enable a transmission in a network between third parties by an intermediary or a lawful use, and which have no independent economic significance.

⁴² Brazilian Copyright Law (Law No. 9,610/1998), Article 29, items III and IV (adaptation, musical arrangement, translation and other transformations). The position of the Ministry of Culture was taken, *inter alia*, in a technical note addressed to the Autoridade Nacional de Proteção de Dados in October 2024 concerning the training of a major developer's models, invoking Article 29(III) (transformation for the making of derivative works); see the report at <https://nucleo.jor.br/curtas/2024/10/10/ministerio-cultura-prejuizos-direitos-ia-meta/> accessed 7 July 2026.

the adaptation right alongside the reproduction right. The position in Canadian, French, and German law is broadly compatible.

(c) Communication to the public and making available

The right of making available to the public, contained in Article 8 of the WIPO Copyright Treaty and in Articles 10 and 14 of the WIPO Performances and Phonograms Treaty, is triggered primarily at the output stage, where the deployed model returns content to the user, if that content contains a copy of protected expression. In Brazil, which is not a Contracting Party to the WCT or WPPT, the *Superior Tribunal de Justiça* held in 2017, in REsp 1.559.264-RJ, that the streaming of music over the internet engages the right of communication to the public, which encompasses what other jurisdictions could identify as the right of making available.⁴³ This is consistent with the “umbrella solution” adopted at the 1996 Diplomatic Conference, under which the making-available right may be implemented through the right of communication to the public, the right of distribution, or a combination of them. The application of analogous reasoning to the outputs of a deployed AI model has not yet been tested in the Brazilian courts.⁴⁴ Performers and producers of phonograms hold parallel rights of making available under Articles 10 and 14 of the WPPT, noted above, so that the analysis extends beyond authors’ rights.

(d) Moral rights

The moral right may also be engaged. The right of integrity is the more frequently invoked. In Brazil, the fragmentation of works during the training process is seen as a potential interference with the integrity right, and the failure to credit the authors of the materials used as a potential interference with the right of attribution.⁴⁵ The Canadian Copyright Act recognizes both the right of integrity and the right of attribution as moral rights of the author.⁴⁶ In a number of *droit d’auteur* jurisdictions, the issue of whether the use of a work in a training corpus without attribution, and in a manner that can generate outputs that are perceived as derogatory or distorted versions of the original, constitutes an interference with moral rights has been raised in academic commentary but has not yet received an authoritative answer.

Finally, in the European Union and in some jurisdictions that have adopted similar protection, the sui generis right in databases protected by Directive 96/9/EC may be engaged where the training corpus is constructed by extracting and re-utilizing substantial parts of a protected database.⁴⁷ The right is independent of copyright and has its own substantial-investment and “substantial-part” criteria (for the “extraction right”). The scope of its application to the construction of AI training corpora remains unsettled, including in relation to legal-information databases, scientific journal aggregators, and large image libraries.

The rights potentially implicated by the training of a generative AI model may therefore extend beyond the reproduction right. Whether a bilateral or collective license clears the

⁴³ Superior Tribunal de Justiça, REsp 1.559.264-RJ (2017) (holding that the streaming of music over the internet constitutes a form of public performance and engages the rights of communication to the public, and that ECAD is entitled to collect on that basis).

⁴⁴ Brazilian Copyright Law (Law No. 9,610/1998), Article 5(V) (definition of “communication to the public”) and Article 29(VII) (distribution by means of any system allowing the user to select the work).

⁴⁵ Brazilian Copyright Law (Law No. 9,610/1998), Article 24, items I-IV. Performers also enjoy moral rights of integrity and authorship over their performances under Article 92.

⁴⁶ Copyright Act, R.S.C. 1985, c. C-42, ss 14.1 and 28.1.

⁴⁷ Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77/20, Article 7 (the sui generis right). See also CJEU, C-203/02, *British Horseracing Board v. William Hill*, ECLI:EU:C:2004:695, on the conditions for the substantial-investment criterion.

relevant uses therefore depends on the rights it covers; a license limited to reproduction may not address claims based on adaptation, making available, moral rights, or, where applicable, the sui generis database right. A further point, taken up in Part III in the context of cross-border enforcement, is that the bundle of rights and the precise scope of each right vary from jurisdiction to jurisdiction, so that a licensing arrangement designed to clear rights in one jurisdiction may leave the developer exposed in another.⁴⁸

(e) RMI and anti-circumvention

Under Articles 11 and 12 of the WCT, Articles 18 and 19 of the WPPT, and Articles 15 and 16 of the Beijing Treaty, rightholders have a right to adequate legal protection and effective legal remedies against the circumvention of effective technological measures that are used by them in connection with the exercise of their rights under those instruments and the Berne Convention, as well as adequate and effective legal remedies against any person knowingly performing the following acts knowing, or with respect to civil remedies having reasonable grounds to know, that it will induce, enable, facilitate or conceal an infringement of any right covered by those Treaty or the Berne Convention:

- (i) the removal or alteration of any electronic rights management information without authority;
- (ii) the distribution, importation for distribution, broadcasting or communication to the public, without authority, works or copies of works, phonograms or performances and (under the respective instruments) knowing that electronic rights management information has been removed or altered without authority.

To our knowledge, no successful suit had been brought as of this writing against companies in respect of the training of AI models.

2. Voluntary (collective and individual) licensing

(a) Individual and collective licensing of AI uses

Voluntary licensing involves the negotiation of an authorization between an AI developer and a rights holder or a representative of many rights holders, including a CMO or other intermediary, without any statutory mandate. This is the default in all jurisdictions studied.

Two structural features shape the form that voluntary licensing takes in practice. The first is the asymmetry between the small number of large institutional rights holders who can negotiate on commercially meaningful terms with developers, and the very sizeable number of individual authors and small rights holders for whom such negotiation is often impracticable due to transaction costs, lack of expertise or other factors. The major bilateral deals concluded over the past three years have been between AI developers and large publishers, news organizations, broadcasters and image libraries; there is no comparable infrastructure for individual authors, photographers, illustrators or musicians, who often depend on the deals concluded by their publishers and other licensees to obtain any share of the value generated.

The scale of some of these arrangements is now substantial. One five-year agreement between a major news publisher and an AI developer was reported at more than

⁴⁸ See *Toronto Star Newspapers Ltd v. OpenAI Inc.*, 2025 ONSC 6217 (Kimmel J., 7 November 2025) (rejecting OpenAI's jurisdictional motion on the basis of a real and substantial connection to Ontario).

US\$250 million,⁴⁹ and the licensing of content for computational use has become a material revenue line for the large image libraries: Shutterstock, Inc.'s full-year 2025 results reported US\$203.3 million in Data, Distribution and Services revenue, representing approximately 20.5 per cent of total revenue of US\$989.9 million; Getty Images separately reported full-year 2025 revenue of US\$981.3 million. The companies announced a merger of equals on 7 January 2025; Getty Images announced on 30 June 2026 that it would terminate the merger agreement after the UK Competition and Markets Authority conditioned clearance on the divestiture of Shutterstock's editorial business. The agreement was reportedly terminated on 7 July 2026.⁵⁰ The selectivity of this market is illustrated by the French press: in March 2024, Le Monde concluded an agreement with OpenAI that it described as the first between a French media outlet and a major generative AI developer.⁵¹ Later that year, APIG sought collective discussions with OpenAI, although the available reporting suggests that those discussions did not lead to a broader licensing arrangement at that time.⁵² A response to this asymmetry has begun to emerge in the form of intermediaries that aggregate the rights of smaller publishers: one business-information provider reports that several thousand of the publishers in its network, drawn from a far larger pool, have opted in to license their content for AI use,⁵³ and a developer of a licensed answer engine reports a comparable aggregation of participating publications.⁵⁴

The second is the chain-of-title issue. In many jurisdictions, where the publisher, broadcaster or repository has been granted certain rights by the underlying author through a licensing or assignment contract entered into before the advent of generative AI, the question of whether

⁴⁹ The five-year agreement between News Corp and OpenAI was reported by The Wall Street Journal at over US\$250 million. The financial terms of such agreements are generally confidential, and the figure is press-reported rather than publicly confirmed by the parties. See "OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over \$250 Million" (The Wall Street Journal, 22 May 2024) <<https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba>> accessed 3 July 2026; "Sam Altman's OpenAI signs content agreement with News Corp" (Reuters, 22 May 2024) <<https://finance.yahoo.com/news/sam-altmans-openai-signs-content-201808782.html>> accessed 3 July 2026.

⁵⁰ Shutterstock, "Shutterstock Reports Full Year 2025 and Fourth Quarter Financial Results" (17 February 2026) <<https://investor.shutterstock.com/node/14551/pdf>> accessed 3 July 2026; Getty Images Holdings, Inc., "Getty Images Reports Fourth Quarter and Full Year 2025 Results" (16 March 2026) <<https://investors.gettyimages.com/node/15726/pdf>> accessed 3 July 2026; Getty Images Holdings, Inc and Shutterstock, Inc., "Getty Images and Shutterstock to Merge, Creating a Premier Visual Content Company" (7 January 2025) <https://www.sec.gov/Archives/edgar/data/1898496/000110465925001732/tm252186d1_ex99-1.htm> accessed 3 July 2026. On the termination of the merger, see Competition and Markets Authority, "Getty abandons Shutterstock merger plans" (July 2026) <<https://www.gov.uk/government/news/getty-abandons-shutterstock-merger-plans>> accessed 31 July 2026.

⁵¹ The agreement, concluded in March 2024, was described by Le Monde as the first between a French media outlet and a major developer of generative AI models. See "Le Monde signs artificial intelligence partnership agreement with OpenAI" (Le Monde, 13 March 2024) <https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html> accessed 3 July 2026.

⁵² The body in question is the Alliance de la presse d'information générale (APIG), which, together with the Syndicat des éditeurs de la presse magazine (SEPM), formally requested the opening of negotiations with OpenAI in June 2024; the developer reportedly responded that it did not see the need to conclude new licensing agreements with French publishers at that stage. See "Intelligence artificielle : un accord de partenariat entre Le Monde et OpenAI" (Le Monde, 13 March 2024) <https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html> accessed 3 July 2026.

⁵³ Dow Jones' Factiva reported that, by the end of 2025, approximately 5,000 publishers, out of a network of around 30,000, had opted to license their content for use in generative AI systems. See Dow Jones, "Factiva Sees Growing Publisher Participation in AI Licensing" (Dow Jones, 2025) <<https://www.dowjones.com>> accessed 3 July 2026.

⁵⁴ ProRata.ai describes its network, in its own words, as comprising "1000+ publications and content creators".

those rights extend to the use of the work in the training of an AI model is both legally uncertain and contractually unprovided for. The uncertainty is not unique to AI: comparable questions arise under agreements between authors and music publishers, between performers and phonogram producers, and between audiovisual producers and the talent engaged for a production, wherever a mode of exploitation was not contemplated at the time of contracting; what is distinctive in the AI context is the scale on which the question arises and the practical difficulty of renegotiation. In China, for example, most authors give their publishers the exclusive right to publish in print and, increasingly, in digital form. They are the rights to reproduce the work in paper form, to distribute paper copies, and to reproduce and distribute digital copies. None of these types of use are naturally interpreted as including use of the work in the training of an AI model. The position of corpora providers and AI developers who obtain copyright licenses from publishers is therefore that the licensed activities are still of uncertain lawfulness in relation to the underlying authors. Even where the developer has obtained a license from the rights holder in the institution, there is the residual risk of action by the underlying author. Depending on the jurisdiction, the exposure is not confined to the licensor: the licensor may be liable for breach of contract or of the law, and the developer, as the party that ultimately benefits from the use, may itself remain exposed to infringement claims by the underlying authors notwithstanding the license. In the audiovisual sector, template licensing agreements are reported to address this risk contractually: the licensor represents and warrants that it holds all rights necessary for the training use. For newspaper articles written by journalists and editors before the relevant amendment to the Chinese Copyright Law took effect on 1 June 2021, the copyright in those articles generally belonged to the journalist or editor and not the news agency. This meant that even a license from the news agency would not have cleared the use of the article for AI training. Constraints of this kind may themselves hinder the emergence of a voluntary licensing market, since a license from an intermediary whose own rights are uncertain adds a further layer of potential infringement to the chain of title. They also have a distributional dimension: where licensing is intermediated by publishers, broadcasters, and repositories, the asymmetry of bargaining power between individual authors or performers and those institutions means that the benefits of licensing may not reach the individuals whose works are used, a consideration relevant to the choice among the licensing models discussed in Section B.

Chinese practice illustrates, however, that even within these structural constraints, voluntary licensing at scale is occurring in sectoral and vertical-domain contexts. Hong Kong radio and television stations partner with AI developers to provide high-quality Cantonese voice data. Digital novels are supplied through cooperation between developers and online publishers commanding large numbers of contracted writers. Images are supplied through cooperation between developers and major picture repositories, which hold them on the basis of transfers or exclusive licenses from the photographers. Scientific papers are supplied through cooperation between developers and academic database operators, which in some cases become co-developers of the resulting domain-specific model. Some corpus providers also produce training materials of their own, contracting photographers, models, and schoolteachers to transfer or license the resulting works for use in training datasets.

Other jurisdictions show similar patterns, though the institutional arrangements are different. In Brazil, the legal regime for voluntary licensing follows the general rules of the Copyright Law. The need for prior and express authorization according to Article 29, and the restrictive interpretation of copyright transactions under Articles 4, 31 and 49 mean that, even if economic rights were completely assigned prior to generative AI, the use of the work for AI training will be disputable, as the contract could not have contemplated specifically a

technology that did not yet exist. Similar doctrines of strict construction are found in French, Italian, and other *droit d'auteur* jurisdictions.⁵⁵

Where the law provides an explicit author's right in respect of stagings, performances, or other defined subject matter, the voluntary licensing landscape may be more straightforward because the underlying entitlements are clearer. Greater complexity may arise for categories of works created by many contributors, such as audiovisual works and phonograms, where uncertainty in the law or in the contracts between producers and individual contributors may either impede voluntary licensing or channel its benefits away from those contributors.

Senegalese law, which expressly recognizes stagings as protected works,⁵⁶ is an example, although the practical application of that recognition to AI training has not yet been tested. In Kenya, the Copyright Act (Cap 130) recognizes the standard bundle of exclusive rights and provides for the collective administration of copyright through CMOs, but the institutional infrastructure has not been designed for computational use, and voluntary licensing for AI training is not yet a documented practice. In Canada, voluntary licensing is the default mechanism and contract negotiation is the central activity. The principal current litigation, *Toronto Star Newspapers Ltd v. OpenAI Inc.*, arose in the context of disputed unauthorized use rather than of licensing per se, and is discussed in Part III of this study.

A recurring feature of voluntary licensing, individual or collective, is the customization of the license terms to the data-consumption profile of the system for which the data are sought, and that profile varies significantly across the categories of generative AI system now in use. Where data is used only once, for initial training or fine tuning of a model whose weights are then fixed, a common structure is a one-time or revenue-share fee for a defined corpus. In situations where the system needs continuous access to the licensed material when generating output, such as RAG and systems that perform ongoing text and data analysis on a library, access is not a single instance, but a continuous relationship. The compensation models that have evolved in practice reflect this, from ongoing fees to charges per crawl or per inference. Apart from price, the main points the parties have to deal with are the extent of use that is allowed, the length of the license, its territorial scope, what happens if it is terminated, and whether the model can be assigned to third parties. The AI context raises the question of whether termination means the developer must stop using a model already trained on the licensed material, retrain it, or destroy stored copies, none of which fit neatly into the remedies developed for earlier licensing relationships. Termination illustrates the difficulty: in traditional licensing, termination restores the rights holder's control over future uses, but a rights holder who terminates a license for AI training cannot, as a practical matter, compel the developer to cease using what the training has already fixed in the model, particularly where the catalogue concerned is small relative to the corpus; the remedy would have to operate, if at all, through retraining or attempts at machine unlearning, whose limits are noted in Part I. Even where such interventions prove feasible, they entail operational and processing costs, which the parties may need to allocate in the agreement itself, adding a further difficulty to the negotiation of such licenses. When such model weights are released or sold to downstream users, the question of transferability arises: Does the contractual obligation attached to a license follow the model weights? A related but distinct question is whether remuneration downstream could be computed from the model itself, for example by reference to the contribution of licensed material to the weights. These questions have not yet been resolved in the contracts which have been made public and are

⁵⁵ On the pre-2021 ownership rule in Chinese copyright law for newspaper articles authored by journalists and editors, see Copyright Law of the People's Republic of China (as amended in 2020, in force from 1 June 2021), Article 18 (works for hire).

⁵⁶ Senegalese Copyright Law (Loi n° 2008-09 du 25 janvier 2008), Article 2. Comparable provisions exist in the OAPI Bangui Accord (2015 Revised Act), Annex VII.

relevant to the possible choice between the remuneration models discussed in Section B.6 below.

At the high-value end, voluntary licensing in the music and audiovisual sectors has mostly taken the form of individualized partnerships that bundle the license with commercial terms going well beyond the use of protected works. In the course of 2025 the major record companies moved from litigation to licensing with the principal developers of AI music tools: two such developers settled the infringement claims against them and agreed to build new, licensed platforms, one settlement being accompanied by the developer's acquisition of a concert-discovery service from the record company, and a streaming service separately agreed with the three major record companies and two further rights holders to develop AI music products on an opt-in basis.⁵⁷ The settlements were preceded by collective statements of principle in the sector: coalitions of industry bodies, artist organizations, and technology companies have articulated principles that condition the use of protected repertoire in AI development on authorization and licensing and call for transparency in training practices.⁵⁸ In the audiovisual sector, a licensing arrangement announced in December 2025, under which a studio was to make more than two hundred of its characters available on a video-generation platform and to take an equity stake of US\$1 billion in the developer, was reportedly abandoned in March 2026 when the developer withdrew the platform.⁵⁹ These arrangements bear in part on the status of outputs, which lies outside the scope of this study; they are noted here only as illustrations of the customized, high-value licensing structures now emerging, and of their fragility.

(b) The main collective licensing models (voluntary, extended, and compulsory)

The most common form of collective licensing is voluntary collective administration. A CMO is granted, by contract, the management of certain rights by its members; it negotiates and concludes licenses within its mandate; the proceeds are distributed to its members. Most jurisdictions with a developed copyright system have voluntary CMOs, but the rights and works they administer differ significantly. Music public-performance societies are the most universal example. In China, five CMOs operate or are in process of full operation: the Music Copyright Society of China, the China Audio-Video Copyright Association, the China Written Works Copyright Society, the Images Copyright Society of China, and the China Film Copyright Association.⁶⁰ The CMOs have publicly expressed their willingness to license for AI training, but as of the date of this study no major collective licensing agreement for AI

⁵⁷ Universal Music Group and Warner Music Group each settled their litigation against Udio in late 2025 and entered licensing arrangements for new platforms; Warner Music Group additionally settled with Suno, an arrangement that included Suno's acquisition of the concert-discovery service Songkick. A further developer, KLAY Vision, concluded agreements with the major record companies and their publishing affiliates. The streaming collaboration was announced by Spotify, with Sony Music Group, Universal Music Group, Warner Music Group, Merlin and Believe, in October 2025.

⁵⁸ The Human Artistry Campaign, formed in 2023 by a coalition of more than 150 organizations, including major music industry bodies and artist representative groups, articulates seven core principles <<https://www.humanartistrycampaign.com>> accessed 25 August 2026; the Principles for Music Creation with AI, introduced by Roland Corporation and Universal Music Group and endorsed by more than fifty organizations <<https://aiformusic.info>> accessed 25 August 2026. Both initiatives are discussed and situated within a broader responsible-design framework in de Berardinis and others, "RAIM: Operationalising Trustworthy AI in Music Generation", cited at Part I, footnote 17, of this study.

⁵⁹ The arrangement between The Walt Disney Company and OpenAI, announced on 11 December 2025, concerned more than 200 characters and a planned equity investment of US\$1 billion; it was said to have been abandoned following OpenAI's withdrawal of its Sora video-generation product on 24 March 2026, no investment having been completed.

⁶⁰ The first three are commonly abbreviated MCSC, AVCA, and CWWCS; the Images Copyright Society of China and the China Film Copyright Association are the two not yet in full operation.

training has been concluded in China.⁶¹ The structural reason, in the Chinese context, is that the CMO's mandate is limited to the rights of its members, and the developer continues to face residual risk in respect of works whose underlying authors are not members.

Similar characteristics can be observed in the Republic of Korea, with the Korea Music Copyright Association and similar bodies. The limited nature of the collective repertoire seems to be only part of the explanation for the failure to conclude licenses. Under the current regulatory regimes, developers may have limited incentives to pursue licenses voluntarily. Two other factors seem to explain why deals are nonetheless being done in the United States. First, the market structure allows a small number of major rights holders and content providers to negotiate directly with developers. Second, the risk of litigation, including statutory damages, provides an incentive to reach an agreement. The settlement-related licensing agreements described in Section B.2 are examples of this. The explanation for why licensing has not yet gained momentum in China and the Republic of Korea must therefore balance these market and legal factors with the limitations of the collective mandate. At the date of this study the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission in the Republic of Korea were involved with the goal of improvement of the regulatory framework.

Institutionally, the United States has a dense network of sectoral collective management bodies. The performing rights organizations (ASCAP, BMI, SESAC, Global Music Rights, and AllTrack) administer public performance rights in musical works; the Mechanical Licensing Collective, established under the Music Modernization Act of 2018, administers blanket mechanical licenses for digital uses of musical works;⁶² SoundExchange administers statutory royalties for the digital performance of sound recordings under section 114 of the Copyright Act; Copyright Clearance Center, Inc. (CCC) administers reproduction rights in published text materials; and the Artists Rights Society (ARS) administers certain reproduction rights in visual art. Each organization operates within its sector and on terms negotiated with its constituency. In response to the emergence of AI training as a use of protected works, CCC has extended its annual corporate license to cover certain AI uses and has announced plans for several additional AI-specific licenses.⁶³

Beyond China and the United States, the past two years have seen several concrete steps by CMOs and other intermediaries to bring AI training within their licensing mandates, several of them announced only in late 2025. In Sweden, STIM launched a collective licensing arrangement in September 2025, involving the AI music company Songfox and the attribution provider Sureel, whose technology traces AI outputs back to the works that influenced them.⁶⁴ In Norway, Kopinor concluded an agreement in December 2025 under which the National Library of Norway may use newspaper content to train Norwegian-language models, funded by the Norwegian State.⁶⁵ In Denmark, the Danish Press

⁶¹ The major Chinese CMOs have publicly expressed willingness to license for AI training, but no significant licensing agreement has been concluded to date.

⁶² Orrin G. Hatch–Bob Goodlatte Music Modernization Act, Pub. L. No. 115-264, 132 Stat. 3676 (2018), codified at 17 U.S.C. §§ 115 and following. The Mechanical Licensing Collective is designated under 17 U.S.C. § 115(d)(3).

⁶³ See www.copyright.com/ai.

⁶⁴ STIM, “STIM launches world’s first collective AI music licence” (September 9, 2025), reported by CISAC, <https://www.cisac.org/Newsroom/society-news/stim-launches-worlds-first-collective-ai-music-licence>. The opt-in licence covers both model training and the downstream use of AI outputs; compensation combines an up-front payment with a share of the developer’s revenue, and the attribution provider Sureel traces outputs back to the works that influenced them.

⁶⁵ “Kopinor signs historic agreement on newspaper content for AI in Norway” (December 19, 2025), reported by IFRRO, <https://ifro.org/page/article-detail/ifro-member-kopinor-signs-historic-agreement-on-newspaper-content->

Publications' Collective Management Organization concluded an opt-in licensing partnership with ProRata AI for the use of members' content in a retrieval-based answer engine, also in December 2025, while a further Danish organization, Tekst&Node, is developing a licensing model under the Danish extended collective licensing framework.⁶⁶ In the Netherlands, the association of private news publishers, together with the news agency ANP, has cooperated with public organizations to develop a domestic language model, GPT-NL, since July 2025.⁶⁷ In Spain, CEDRO has joined the Really Simple Licensing initiative to develop an AI licensing model for the Spanish market.⁶⁸ In the United Kingdom, the Copyright Licensing Agency, with Publishers' Licensing Services and the Authors' Licensing and Collecting Society, has been developing a dedicated collective license for generative AI, covering training, fine-tuning, and RAG.⁶⁹ Elsewhere, reproduction-rights organizations have incorporated AI re-use rights into their existing digital copyright licenses or are developing comparable solutions: in Japan, JAC added such rights to its Digital Copyright License in April 2025;⁷⁰ the Copyright Agency in Australia has been developing comparable solutions;⁷¹ and in

for-ai-in-norway/. Concluded by Kopinor on behalf of the Norwegian Media Businesses' Association, the agreement grants the National Library of Norway access to newspaper content older than one year for the training of Norwegian-language models and is funded by the Norwegian State at NOK 45 million per year. It is distinct from the earlier Bookshelf agreement on the digital dissemination of books.

⁶⁶ ProRata AI and the Danish Press Publications' Collective Management Organisation (DPCMO), "ProRata Partners with Danish Publishers Group DPCMO" (December 10, 2025), <https://www.prnewswire.com/news-releases/prorata-partners-with-danish-publishers-group-dpcmo-to-launch-the-first-decentralized-sovereign-ai-answer-engine-302637182.html>. DPCMO, which states that it represents 99 per cent of the Danish news industry, entered an opt-in collective licensing partnership enabling members' content to be used in ProRata's retrieval-based answer engine on a revenue-sharing basis. A further Danish organization (Tekst&Node) is developing an AI licensing model under the Danish extended collective licensing framework, which allows rights holders to license certain text and data mining and AI uses of literary works, subject to opt-outs and approval by the Ministry of Culture. See Terese Foged, "[DK] First extended collective licence approval for text and data mining and the use of works by AI" (IRIS Legal Observations of the European Audiovisual Observatory, 2026) <<https://merlin.obs.coe.int/article/10505>> accessed 3 July 2026.

⁶⁷ NDP Nieuwsmedia and TNO, "Large dataset news organizations for Dutch AI language model GPT-NL" (July 17, 2025), <https://www.ndpnieuwsmedia.nl/2025/07/17/large-dataset-news-organizations-for-dutch-ai-language-model-gpt-nl/>. Members of NDP Nieuwsmedia, the association of Dutch private news publishers, together with the news agency ANP, are making portions of their archives available to train GPT-NL, a non-profit Dutch language model developed by TNO, NFI, and SURF; contributing publishers are to receive remuneration on the model's release.

⁶⁸ CEDRO, press release on its collaboration agreement with the RSL Collective (January 2026), <https://www.cedro.org/sala-de-prensa/noticias>. CEDRO, the Spanish reprographic-rights organization, has joined the RSL Collective as its partner in Spain and participates in the RSL technical committee, adapting the machine-readable licensing standard to the Spanish market so that its members may set the terms on which their works are used in AI training. CEDRO separately published recommendations on the assignment of rights for AI training on December 16, 2025.

⁶⁹ Copyright Licensing Agency, "CLA announces development of Generative AI Training Licence" (2025), <https://cla.co.uk/development-of-cla-generative-ai-licence/>; Publishers' Licensing Services, "PLS and ALCS agree to development of pioneering CLA generative AI licence" (April 2025), <https://www.pls.org.uk/news-events-policy/news/pls-and-alcs-agree-to-development-of-pioneering-cla-generative-ai-licence/>. The CLA, with PLS and ALCS, is developing an opt-in collective licence for the use of text in generative AI, covering training, fine-tuning, and retrieval-augmented generation; rights holders participate by registering their works.

⁷⁰ Japan Academic Association for Copyright Clearance (JAC) and RightsDirect Japan, "JAC and RightsDirect Japan Announce the Availability of AI Re-Use Rights for Digital Copyright License" (April 9, 2025), <https://www.businesswire.com/news/home/20250409032666/en/>. JAC added AI re-use rights to its Digital Copyright License, an enterprise-wide internal-use license for companies headquartered in Japan, drawing on a repertoire of works from publishers and academic societies worldwide.

⁷¹ Copyright Agency, "AI and Copyright in Australia" and accompanying Q&A, <https://www.copyright.com.au/membership/ai-and-copyright-in-australia/>. The Copyright Agency, a member of the Australian Government's Copyright and AI Reference Group, states that it is developing licensing solutions for the use of protected material in connection with generative AI in Australia.

Germany, VG Wort has extended its Digital Copyright License to AI re-use, while the collecting society GEMA has separately proposed a dedicated model that had not yet produced a concluded agreement at the date of this study.⁷² These developments remain uneven and, in several cases, experimental, but they indicate that the collective-licensing route is being actively tested rather than merely debated.

Extended collective licensing (ECL) is a licensing model that originated in the Nordic jurisdictions in the 1960s and has been adopted in several other countries. It addresses the residual-risk problem by extending the effect of a CMO's license to non-member rights holders within a defined repertoire and provides a right of opt-out. The model has been mentioned in European policy debates on AI training as a means of clearing the rights in large textual and visual corpora that the developer needs to use but for which individual licensing is not practical. A 2025 European Parliament Study on Generative AI and Copyright explicitly mentioned it.⁷³ In the United States, the Copyright Office in the third part of its Report on Copyright and Artificial Intelligence identified extended collective licensing as a possible targeted intermediate option for AI training, distinct from compulsory licensing.⁷⁴ The same proposal is reflected, with variations, in the academic literature on remuneration: Professor Senftleben's work on Generative AI and Author Remuneration proposes a system in which collective management organizations would collect remuneration for the use of generative AI services whose outputs substitute for or compete with the works of identifiable authors, and would distribute the proceeds among their members.⁷⁵

In addition to the examples already mentioned, two other examples of how existing, voluntary collective management could be adapted for AI training, working within the existing institutional structures rather than waiting for new legislation, illustrate this possibility. In Brazil, the public-performance regime of music is centralized in one office, ECAD, which has legal and judicial backing to aggregate the collection and distribution of public-performance royalties on musical works and phonograms, for authors' as well as related rights.⁷⁶ Working within the existing Copyright Law and absent legislative change, it would be legally possible for a specific right regarding the use of works and phonograms for the training of AI models to be exercised collectively under a model similar to the one currently in place for the right of public performance of music. This would require that the rights holders join a CMO and that the CMO provide in its bylaws for the representation of the rights holder in relation to other copyrights, and not only for the rights of public performance of music; and it would require prior authorization from the Ministry of Culture, the federal government agency responsible

⁷² On VG Wort, the Digital Copyright License administered through RightsDirect (a CCC subsidiary) has been extended to internal AI re-use; see <https://www.rightsdirect.com/license/vg-wort-digital-license-2/>. The collecting society GEMA proposed a licensing model for generative AI in 2024; no agreement had been concluded at the date of this study, and GEMA has in parallel brought infringement proceedings against AI developers (see *GEMA v. OpenAI*, *supra*). See GEMA, "AI and music: GEMA calls for music creators to share in AI providers' revenues and presents first licensing model" (26 September 2024) <<https://www.cisac.org/Newsroom/society-news/ai-and-music-gema-calls-music-creators-share-ai-providers-revenues-and>> accessed 3 July 2026; GEMA, "Suno AI and Open AI: GEMA sues for fair compensation" (20 January 2025) <<https://www.gema.de/en/news/ai-and-music/ai-lawsuit>> accessed 3 July 2026.

⁷³ European Parliament, Directorate-General for Internal Policies, Study on "Generative AI and Copyright: Training, Creation, Regulation", IUST_STU(2025)774095_EN (2025), recommending statutory licensing as the optimal framework for AI training in the EU.

⁷⁴ United States Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (2025).

⁷⁵ M.R.F. Senftleben, "Generative AI and Author Remuneration", 54 IIC 1535 (2023).

⁷⁶ On the structure of the Brazilian collective management system and the centralization of public-performance royalty collection in ECAD, see Brazilian Copyright Law (Law No. 9,610/1998), Articles 97-100, and Decree No. 9,574/2018. The constitutionality of the model was confirmed by the Supremo Tribunal Federal in ADI 2.054 (2003) and in the joint ruling on ADIs 5062 and 5065 (2016).

for CMO oversight. The Brazilian Copyright Law, in Article 97, allows CMOs to expand their scope to the licensing of the rights implicated by training and by the implementation of the model (the reproduction right, communication to the public or making available and, on the position taken by the Ministry of Culture, the adaptation right in the output phase), provided the requisite authorizations are obtained. The unitary-collection requirement, however, applies only to music public-performance rights, so that even if ECAD or its affiliates were to centralize the representation of an AI training right within the body, this would not preclude the formation of other collective entities with the same purpose.

In Kenya, the Copyright (Collective Management) Regulations 2020 govern the registration, operation, and supervision of CMOs.⁷⁷ The framework is general and not specifically adapted to computational uses, but it provides an institutional basis on which voluntary collective licensing for AI training might be built. The Draft Copyright and Related Rights Bill 2026, published for public comment by the Kenya Copyright Board in March 2026, with the comment period closing at the end of that month, contains a section (Section 34) permitting text and data analysis for non-commercial research, but the bill does not provide a clear statutory definition of “non-commercial”, does not introduce an opt-out mechanism for rights holders, and does not contain specific provisions either on the use of copyrighted works in AI training or on the copyrightability of AI-generated works. The Kenyan situation is illustrative of a wider pattern in jurisdictions where the copyright framework is being adapted in stages: the existing CMO infrastructure provides a possible vehicle for collective licensing of AI training, but the adaptation requires positive legislative or contractual steps that have not yet been taken.

3. Community and stewardship licensing of primary language data

The licensing mechanisms described above operate on training data that is already fixed and published when the developer encounters it. A distinct category of training data comes into existence differently. Speech and language models for most African languages cannot be built from web-scale text corpora, because little such text exists. The data is instead collected as primary data: audio recordings and transcriptions gathered from speakers through fieldwork, community recording sessions, and partnerships with broadcasters. Copyright attaches at the point of fixation rather than at the point of crawling. The recording is protected as a sound recording, the transcription as a literary work, and the speaker may hold performers’ rights.⁷⁸ Questions of authorization and compensation therefore arise at the moment of collection, before any question of text and data mining can occur.

This category of data also engages a further layer of intellectual property. Recordings of oral tradition, folktales, and proverbs are at once copyright works vesting in the person who fixed them and cultural expressions in which communities hold collective rights under *sui generis* legislation. In Kenya, the Protection of Traditional Knowledge and Cultural Expressions Act, No. 33 of 2016, which gives effect to Articles 11, 40, and 69(1)(c) of the Constitution of Kenya, 2010, recognizes communities as rights holders in their cultural expressions, requires prior informed consent and authorized user agreements for their exploitation, and provides for compensation flowing to communities; Article 11(3)(a) of the Constitution obliges Parliament to ensure that communities receive royalties for the use of their cultures

⁷⁷ The Copyright (Collective Management) Regulations 2020 (Kenya). The licensing infrastructure is set out in Sections 33 and 34 and Part VII of the Copyright Act (Cap 130).

⁷⁸ Copyright Act (Cap. 130), Laws of Kenya, section 2 (defining, *inter alia*, “sound recording” and “literary work”) and the provisions on the rights of performers. On performers’ rights at the international level, see Articles 10 and 14 of the WPPT, discussed in Section B.1 above.

and cultural heritage.⁷⁹ Where language data of this kind enters a training corpus, the copyright chain of title and the community entitlement operate in parallel, and a license from the fixer of the recording does not of itself address the statutory position of the community.

A distinct licensing practice has developed in response. Standard open licenses, and in particular public domain dedications, place such data in training pipelines without providing governance authority or economic return to the contributing community. Community-centered licenses modify this position while retaining openness. The Nwulite Obodo Open Data Licence applies tiered conditions of use differentiated by the user's geographic and development context and incorporates benefit-sharing obligations; it has been applied to African speech datasets, including the DhoNam Dholuo Speech Dataset developed with Maseno University.⁸⁰ The Esethu License, developed for the Vuk'uzenzele isiXhosa Speech Dataset, permits research use, and commercial use by African entities, without charge, while requiring licensing fees from other commercial users, the proceeds being reinvested in the creation of further language data.⁸¹ Te Hiku Media's Kaitiakitanga License, governing Māori language data, reflects the same approach in another region.⁸² These instruments constitute a third mechanism, alongside individual and collective licensing, through which authorization and compensation for the use of primary language data in AI training are being organized, and they illustrate how a national statutory framework for collectively held cultural assets can underwrite contractual practice. In Kenya, this layer will also interact with the text and data analysis exception proposed in the Draft Copyright and Related Rights Bill 2026, discussed in the preceding subsection.

4. Statutory TDM exceptions and general clauses

European Union copyright law contains explicit text and data mining exceptions relevant to operations used in AI training, although their application and adequacy in that context remain contested. Articles 3 and 4 of Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market (the CDSM Directive) establish two mandatory exceptions for text and data mining. Article 3 is a research-purpose exception for research organizations and cultural heritage institutions, without an opt-out. Article 4 is a general exception applicable to any beneficiary, subject to an opt-out by the rights holder "in an appropriate manner, such as machine-readable means in the case of content made publicly available online".⁸³ The AI Act incorporates the framework by reference, obliging providers of general-purpose AI models to put in place a policy to comply with Union copyright law and to identify and respect the

⁷⁹ Protection of Traditional Knowledge and Cultural Expressions Act, No. 33 of 2016 (Kenya); Constitution of Kenya, 2010, Articles 11, 40(5), and 69(1)(c).

⁸⁰ Nwulite Obodo Open Data Licence, Version 1.0 (Data Science Law Lab, University of Pretoria, 2025) <<https://datasciencelawlab.africa/nwulite-obodo-open-data-license/>> accessed 31 July 2026; C.I. Okorie and M. Omino, "Addressing Inequitable Openness in Licences for Sharing African Data and Datasets Through the Nwulite Obodo Open Data Licence" (2025) 7(3) Law, Technology and Humans 94. The DhoNam Dholuo Speech Dataset (Maseno University Centre for African AI) is released under that license: <<https://mozilladatacollective.com/datasets/cmjepxo6t08nmmk07iauvua6v>> accessed 31 July 2026.

⁸¹ Lelapa AI, Way With Words, and Data Science for Social Impact (University of Pretoria), "The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (2025) <<https://aclanthology.org/2025.acl-long.1487/>> accessed 31 July 2026.

⁸² Te Hiku Media, Kaitiakitanga License <<https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license>> accessed 31 July 2026.

⁸³ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, OJ L 130/92 (the "CDSM Directive"), Article 3 (mandatory exception for text and data mining for the purposes of scientific research by research organizations and cultural heritage institutions) and Article 4 (general, opt-out exception for text and data mining of lawfully accessible works).

reservations of rights expressed pursuant to Article 4(3) of the CDSM Directive, and to publish a sufficiently detailed summary of the content used for training.⁸⁴

Two German rulings have addressed the application of those exceptions to AI training. In *Kneschke v. LAION e.V.*, the Landgericht Hamburg held in September 2024 that the use of a photograph in the construction of the LAION-5B dataset, which is used in the training of image-generation models, was covered by the German implementation of the Article 3 CDSM exception for text and data mining for scientific research. The court treated LAION as a scientific research organization within the meaning of the provision and considered that the computational analysis at issue constituted text and data mining for scientific purposes. It also indicated, obiter, that the Article 4 general TDM exception would in any event not apply, because the rights reservation contained in the source platform's terms of use was, in the court's view, machine-readable in light of contemporary large language model capabilities.⁸⁵ On appeal, in *Kneschke v. LAION e.V.*, the Oberlandesgericht Hamburg confirmed the outcome but provided a more developed treatment of the machine-readability requirement under Article 4(3).⁸⁶ The court of appeal held that machine-readability requires not only that the reservation be capable of being captured by machine, but also that it be capable of being interpreted by an automated process in such a way that the content covered by the reservation is not processed. The reservation must, in short, be both machine-readable and machine-actionable. The court further indicated that machine-readability is not a static property but must be assessed against the technological capabilities prevailing at the time of the disputed use; on the facts of the case, that meant the state of the art in the second half of 2021. The court of appeal's reasoning leaves open the question of what would constitute machine-readability in 2025 or later, when models capable of extracting meaning from unstructured natural-language text are widely available.⁸⁷

About a month earlier, in *GEMA v. OpenAI*, the Landgericht München I had reached a different conclusion on a related but distinct question.⁸⁸ The German collecting society GEMA had sued OpenAI alleging that ChatGPT had memorized, and could be induced to reproduce, the lyrics of nine well-known German songs. The Munich court treated both the ingestion of the lyrics into the training corpus and their subsequent memorization within the model as acts of reproduction within the meaning of the German implementation of the InfoSoc Directive; it rejected the application of the general TDM exception in § 44b of the German Copyright Act on the theory that generative models permanently memorize lyrics and that this goes beyond the analysis of textual and pattern-based information that the

⁸⁴ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, Article 53(1)(c) and (d), and Recitals 105 and 106.

⁸⁵ *Kneschke v. LAION e.V.*, LG Hamburg, 27 September 2024 (310 O 227/23) (holding that the construction of the LAION-5B dataset for use in training image-generation models was covered by the German implementation of the Article 3 CDSM exception for text and data mining for scientific research (§ 60d UrhG); also indicating, obiter, that the Article 4 general TDM exception would not have applied, on the basis that the rights reservation contained in the source platform's terms was machine-readable in the relevant sense). On appeal, see *Kneschke v. LAION e.V.*, OLG Hamburg, 10 December 2025 (5 U 104/24), discussed below.

⁸⁶ *Kneschke v. LAION e.V.*, OLG Hamburg, 10 December 2025 (5 U 104/24).

⁸⁷ The Hamburg Court of Appeal stated that machine-readability requires that the reservation "can be interpreted by a machine in such a way ... that in an automated process, the content covered by the reservation [of rights] is not processed" (authors' translation). See also P. Keller, "LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs", Kluwer Copyright Blog (17 and 18 December 2025), Part 1.

⁸⁸ *GEMA v. OpenAI*, LG München I, 11 November 2025 (42 O 14139/24), Endurteil. The plaintiff is the German collecting society GEMA; the defendants are OpenAI OpCo LLC and OpenAI Ireland Limited. The court ordered injunctive relief, the provision of information, and damages.

exception was designed to permit; and it read the regurgitated outputs as proof that the lyrics were fixed inside the model in a manner comparable to a digital library. The court granted injunctive relief, ordered the provision of information, and awarded damages.⁸⁹ The decision is under appeal. It has been read in some quarters as the first major European AI-training judgment in favor of rights holders; others have criticized it as collapsing into a single act of reproduction what is, technologically, a sequence of distinct operations, and as misapplying the TDM exception by conflating evidence of memorization at the output stage with the technical reality of the training process.⁹⁰

Beyond the case law, there is an academic dispute about whether the Article 4 CDSM exception applies to AI training at all. A widely held view in the literature is that AI training is a form of data mining covered by Articles 3 and 4 CDSM. This view often cites the fact that those articles are referenced in the AI Act. A contrary view has been articulated prominently by a number of EU scholars, including Professor Tim Dornis, who argues, in a paper published in the *European Intellectual Property Review* and in a more extensive legal opinion co-authored with the computer scientist Stober, that a close examination of the technological foundations of generative AI training reveals that the operations performed go beyond what the text and data mining exception was designed to cover. The argument runs broadly as follows: the TDM exception permits the extraction of semantic, factual, and statistical information from works; the training of a generative AI model, by contrast, extracts and encodes syntactic and expressive features as well, in a manner that is then used to produce output substituting for the original works; and the resulting interference with the legitimate interests of rights holders exceeds the scope of what the exception was framed to permit.⁹¹

Professor Eleonora Rosati has developed a similar argument but from a different starting point. She takes the position that text and data mining may be part of AI training processes but is neither synonymous with AI training nor is it all that AI training entails; that limiting the European debate to Article 4 of the CDSM Directive is too narrow, given the multiplicity of acts restricted by copyright that AI training may involve; and that, in any event, the three-step test in Article 7(2) of the CDSM Directive and Article 5(5) of the InfoSoc Directive may operate to exclude AI training from the exception, on the ground that the use conflicts with a normal exploitation of the works and unreasonably prejudices the legitimate interests of the rights holders. If this is correct, licensing remains a necessary part of the analysis.⁹² Professor Matthias Leistner has separately argued that the opt-out mechanism in Article 4(3)

⁸⁹ Ibid. The court treated both the ingestion of the lyrics into the training corpus and their subsequent memorization within the model as acts of reproduction within the meaning of § 16 UrhG, implementing Article 2 of Directive 2001/29/EC.

⁹⁰ For a critical commentary on this point, see G. Frosio, "Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI", *Kluwer Copyright Blog* (10 and 11 December 2025), Parts 1 and 2, arguing that the Munich court conflates evidence of memorization at the output stage with the technical reality of training, and that the better view is to treat lawful intermediate technical copies as falling within Article 5(1) InfoSoc or the TDM exceptions.

⁹¹ T.W. Dornis, "The Training of Generative AI Is Not Text and Data Mining", 47 *European Intellectual Property Review* 65 (2025). See also T.W. Dornis and S. Stober, *Urheberrecht und Training generativer KI-Modelle - technologische und juristische Grundlagen*, Initiative Urheberrecht (Nomos, September 2024).

⁹² E. Rosati, "Is text and data mining synonymous with AI training?", 19 *Journal of Intellectual Property Law and Practice* 851 (2024); E. Rosati, "The German LAION decision: A problematic understanding of the scope of the TDM copyright exceptions and the transition from TDM to AI training", *IPKat* (7 October 2024); E. Rosati, Copyright Exceptions and Fair Use Defences for AI Training Done for 'Research' and 'Learning', or the Inescapable Licensing Horizon, SSRN Working Paper No. 5331405 (2025).

is likely to become the central challenge in EU copyright law and is at present unworkable in practice, a position that has been taken up in the European policy debate.⁹³

Professor Senftleben argues instead that article 4 Is not only applicable in the context of AI training but also that it may be directly relevant in determining what constitutes normal exploitation of a work in the AI context, which would be relevant in the application of the three-step test.⁹⁴

The European Commission, in its interpretation of the AI Act, has treated Article 4 CDSM as applicable to AI training in the sense that their approach focuses on implementation of the opt-out mechanism rather than on excluding AI training from the scope of the exception.. The opt-out infrastructure is taken up in Part III of this study, in the context of transparency and enforcement.⁹⁵

The Korean legislature has considered but not enacted a text and data mining exception; bills to that effect have been before the National Assembly since 2021.⁹⁶ In the absence of a statutory provision, the question of training is to be evaluated under the general fair use clause in Article 35-5 of the Copyright Act, a provision modeled on section 107 of the United States Copyright Act and introduced following the free trade agreement between the Republic of Korea and the United States. The domestic debate closely follows the considerations surveyed in this Section. Proponents of a statutory exception point to the certainty that would be provided by express statutory language; those opposing a statutory exception respond that statutory language creates its own interpretive ambiguities, as the German litigation over machine readability discussed above shows, that the 2024 articulation by the Supreme Court of Korea of the criteria governing Article 35-5 gives the general clause workable content, and that a broad exemption could disproportionately benefit foreign developers while transferring value away from the holders of rights in Korean-language content.⁹⁷

When dealing with training, the general clause has been given operational content through administrative guidance. In February 2026, the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission jointly issued the Guide on Applicability of the Fair Use Doctrine to Training of Generative AI Models (referred to in Part I of this study), which applies the four factors of Article 35-5 to the stages of the training pipeline described in Part I of this study. The Guide explains that the manner in which the training material was obtained is relevant, so that unauthorized scraping, bypassing technological protection measures, circumventing paywalls, and disregarding reservations expressed through the “robots” exclusion protocol weigh against a finding of fair use, and illustrative examples are provided on either side of the line.⁹⁸

⁹³ M. Leistner, “TDM und KI-Training in der Europäischen Union”, GRUR 2024, 1665 (arguing that the opt-out mechanism is likely to become the central challenge in EU copyright law and is at present unworkable in practice).

⁹⁴ Senftleben, M. Text and Data Mining, Generative AI, and the Copyright Three-Step Test. IIC 57, 67, at 79 (2026). <https://doi.org/10.1007/s40319-026-01680-2>

⁹⁵ AI Act, Article 53(1)(c) (obligation to put in place a policy to comply with Union copyright law and identify and respect, including through state-of-the-art technologies, the reservations of rights expressed pursuant to Article 4(3) CDSM).

⁹⁶ Beginning with the comprehensive Copyright Act amendment bill introduced in the National Assembly on 15 January 2021, which contained a text and data mining exception; the bill lapsed with the twenty-first National Assembly, and successor proposals remain under consideration.

⁹⁷ Supreme Court of Korea, Decision 2021Da272001 (2024), the governing articulation of the criteria under Article 35-5 of the Copyright Act.

⁹⁸ Korea Fair Use Guide, above note 10.

5. Remuneration systems

(a) Types of remuneration systems

Where the use of protected works in AI training is permitted, whether by an exception, a license, or both, the question of remuneration may arise. Several models are in play, drawing on the broader experience of copyright law in adapting to new exploitation patterns.

The first is a system of licensing fees, negotiated either individually or collectively, paid by the developer at the point of corpus assembly or training. The strength of the model is that the rights holder is paid for the use of the work in training, without further qualification; the weakness in the case of individual licensing is that the model leaves out the long tail of individual rights holders for whom direct negotiation is not feasible.

The second is a levy-based system, in which a charge is imposed on AI developers or on certain categories of users of generative AI services, and the proceeds are distributed to rights holders through a CMO or comparable body. The European Parliament Study referenced above in Section B.3 recommends a form of statutory licensing for AI training in the European Union, which would have features of a levy-based scheme combined with mandatory collective administration. Comparable proposals have been made for jurisdictions outside the European Union, often by analogy to the private-copying levies that exist in many countries. Questions have arisen concerning the compatibility of such schemes with the three-step test in Article 9(2) of the Berne Convention, Article 10 of the WCT, and Article 13 of the TRIPS Agreement, and, to the extent rights other than reproduction are swept in, with the specific conditions of Articles 11*bis*(2) and 13(1) of the Berne Convention.

The third, most recent, and conceptually most distinct model is output-level remuneration. On this approach, the focus of the remuneration is not the training use itself but the production of outputs by the deployed model that substitute for or compete with the works of identifiable authors. The model links remuneration directly to possible substitution of the rights holder's work in the relevant market. Its implementation, however, would seem to require technical mechanisms to identify substituting outputs and trace them back to identifiable works in the training data.

Two further issues arise. Where the number of identified works is very large, the per-work remuneration may be small to the point of being nominal. And the basis of calculation requires clarification: in most proposals, remuneration would rest not on a showing that a given output reproduces or adapts a specific work, but on reporting by the developer of the ways in which works were used in the system, an approach that presupposes the transparency mechanisms discussed in Section C of this Part.⁹⁹ Moreover, authors may be affected by substitution whether or not their work was used for training. As noted above, with current filtering technologies, in practice outputs are unlikely to infringe any particular work protected by copyright, so that the harm considered is systemic rather than typical copyright infringement. A reproduction made at the training stage, or subsisting in the model, does not entail that outputs infringe: the infringement analysis for any given output turns on the resemblance of that output to a protected work.

The scale of the substitution concern is increasingly documented in the music sector: in April 2026, Deezer reported that fully AI-generated tracks accounted for approximately 44 per cent of its daily uploads, or around 75,000 tracks per day, while Spotify reported the removal

⁹⁹ In the music domain, the integrity of the work-level and rights-holder-level metadata attached to training material has been identified as the design precondition both for attribution and for any sharing of revenues derived from generated output: de Berardinis and others, "RAIM: Operationalising Trustworthy AI in Music Generation", cited at Part I, footnote 17, of this study.

of more than 75 million spam tracks over the preceding 12 months,¹⁰⁰ and an industry study has projected that a substantial proportion of music creators' revenue could be displaced by AI-generated content by 2028.¹⁰¹ In response, Professor Senftleben has argued that the substitution test could serve as the conceptual vehicle for distinguishing the legitimate use of AI to support human creativity, which would attract no remuneration, from the use of AI to substitute for human creativity, which would. A similar approach arguably undergirds the Senate version of the Brazilian Bill 2338/2023, which couples an opt-out mechanism with a right to remuneration that the rights holder may exercise individually or collectively.¹⁰²

. *(b) Modalities and basis of remuneration*

In Brazil, the substitution effect of synthetic creations on human works is observable in concrete sectors, including ambient sound design, where users have sought to play music generated entirely by AI systems with the aim of evading public-performance fees. The cases pending against ECAD on this question will, in due course, produce judicial guidance on whether the public performance of AI-generated music in commercial premises triggers any remuneration obligation. The example illustrates the interaction between remuneration for AI training and remuneration associated with AI outputs.

Two questions about a possible remuneration system have emerged in ongoing discussions. The first is whether the remuneration should attach to the training), the substitution effect (that is, to, or both. The European Parliament Study and several academic proposals attempt to combine the two; the Brazilian Bill 2338/2023, in its current form, is better read as attaching compensation to ingestion for training: Article 65 provides for compensation in respect of the use of protected content in data mining and in the training and development of AI systems, so that the Bill establishes a compensation mechanism attached to the act of training not extending to the exploitation of outputs.¹⁰³ The second question is whether the remuneration system presupposes a recognized exclusive right that has been licensed or waived, or whether it should be conceived instead as a *sui generis* claim for equitable remuneration, perhaps modelled on the private-copying remuneration that exists in many jurisdictions.¹⁰⁴ As noted above, existing Brazilian legislation might reach the output phase through the right of reproduction, the right of adaptation or the right of communication to the public even if Bill 2338/2023 is not approved, and on that basis the prior and express authorization of rights holders would be required.¹⁰⁵ Private copying, in contrast, is a form of

¹⁰⁰ Deezer, "AI-generated tracks now represent 44% of all new uploaded music" (19 April 2026) <<https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>> accessed 3 July 2026; Spotify, "Spotify Strengthens AI Protections for Artists, Songwriters, and Producers" (25 September 2025) <<https://newsroom.spotify.com/2025-09-25/spotify-strengthens-ai-protections/>> accessed 3 July 2026.

¹⁰¹ International Confederation of Societies of Authors and Composers (CISAC), study by PMP Strategy (December 2024), projecting that some 24 per cent of music creators' revenue could be at risk by 2028 as a result of generative AI, with a higher figure for digital revenue specifically.

¹⁰² Senate Bill No. 2338/2023, as approved by the Brazilian Federal Senate on 10 December 2024 and transmitted to the Chamber of Deputies. The bill is, at the date of this study, before a Special Committee of the Chamber of Deputies; the version that may eventually be submitted to presidential approval is not yet final.

¹⁰³ This is the reading of Article 65 of Bill 2338/2023, as approved by the Federal Senate, adopted in this study. The Bill has also been read as combining an ingestion basis with sensitivity to output substitution, notably because the criteria bearing on compensation refer to market effects. On the reading adopted in the text, the compensation attaches to data mining, training, and development.

¹⁰⁴ A full discussion of possible *sui generis* regimes and which uses would be covered is beyond the scope of this study.

¹⁰⁵ This is the reading of Article 65 of Bill 2338/2023, as approved by the Federal Senate, adopted in this study. The Bill has also been read as combining an ingestion basis with sensitivity to output substitution, notably

compensation, not a license, because the private acts of copying are seen as inherently unlicensable for various reasons, including users' privacy.

In Brazil, a further observation bears on remuneration design: the substitution effect of synthetic creations on human works is observable in concrete sectors, including ambient sound design, where users have sought to play music that was entirely generated by AI systems with the aim of evading public-performance fees. The cases pending against ECAD on this question will, in due course, produce judicial guidance on whether the public performance of AI-generated music in commercial premises triggers any obligation to remunerate ECAD members. The example illustrates the interaction between remuneration for AI training and remuneration associated with AI outputs; choices in one area may affect the design of the other.¹⁰⁶

Finally, in jurisdictions where the regulatory framework is still in formation, the remuneration question is at present open. The 2026 Guide on the Applicability of the Fair Use Doctrine to the Training of Generative AI Models, issued jointly by the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission and referred to in Part I, is principally addressed to the question of what training operations may be undertaken without a license; it leaves the corresponding question of remuneration to be addressed in the broader reform process now under way.¹⁰⁷ Much of the domestic debate, however, proceeds from the premise that commercial training of AI models should carry compensation for creators, so that the question that remains open is the mechanism of remuneration. The Kenyan, Brazilian, and Chinese situations are, in their different ways, comparable: in each, the legislative process is engaged, the policy options are open, and the model that will emerge is not yet determined. Part III of this study, addressing enforcement, returns to the practical consequences of this state of legal flux for rights holders seeking to vindicate their claims.

PART III. ENFORCEMENT

The technical operations described in Part I often engage exclusive rights recognized by national copyright legislation and by the international copyright treaties; the licensing arrangements canvassed in Part II offer mechanisms by which those rights may be cleared when no exception or limitation applies. Part III addresses what happens when neither has occurred: when a rights holder considers that his or her work has been used in the training of a generative AI model without authorization and without the coverage of an applicable exception, and seeks to vindicate his or her rights through the courts or through administrative procedures.

Until recently there was very little case law to consider. That is no longer the case. As of 22 August 2026, 137 suits relating to the use of protected works by generative AI systems had been filed against AI developers in the United States, with approximately 30 further suits recorded in other jurisdictions.¹⁰⁸ Courts in Brazil, Canada, China, Germany, the United

because the criteria bearing on compensation refer to market effects. On the reading adopted in the text, the compensation attaches to data mining, training, and development.

¹⁰⁶ Bill No. 2338/2023, Articles 62 and following (in the version transmitted to the Chamber of Deputies). The trade-secret carve-out qualifies the transparency obligation and is the subject of substantial debate in the legislative process.

¹⁰⁷ Ministry of Culture, Sports and Tourism and Korea Copyright Commission, *Guide on Applicability of the Fair Use Doctrine to Training of Generative AI Models* (February 2026); see Part I, footnote 10, of this study.

¹⁰⁸ According to the litigation tracker maintained by Professor Edward Lee at chatgptiseatingtheworld.com, the United States map showed 156 suits as of 22 August 2026, with approximately 30 further suits recorded in other

Kingdom, the United States and elsewhere have issued rulings that bear on multiple aspects of the analysis: on whether the trained model is or contains copies of the works on which it was trained; on whether exceptions in European law extend to AI training; on whether national courts have jurisdiction over training conducted abroad; on whether outputs that substitute for the works in the training corpus engage the rights holder's right of reproduction or right of making available; and on whether outputs are themselves copyrightable (a question outside the scope of this study, noted here only as context).

The Part is organized around three principal questions. Section A draws lessons from the body of case law as it currently stands, organized by the issue addressed rather than by jurisdiction. Section B turns to the structural question of enforcement against AI developers, including the cross-border features that complicate enforcement. Section C identifies the principal practical challenges that confront rights holders seeking to enforce their rights, including transparency, the proliferation of technical standards for opt-outs, and the operational difficulty of evidence-gathering.

SECTION A. LESSONS FROM CURRENT CASE LAW

The case law that has accumulated over the past eighteen months is fragmentary, jurisdictionally varied, and not yet doctrinally settled in any single legal system. It has not yet addressed all the questions. It nevertheless illuminates at least four distinct issues, on each of which courts in different jurisdictions have begun to converge or, in some cases, to diverge. Those issues are addressed in turn below.

1. The model-as-copy question

The first question is whether, as a matter of copyright law, the trained AI model itself contains copies of the works on which it was trained. This is distinct from the anterior question whether reproductions are made during corpus construction and training. The technical nature of this question of memorization was discussed in Part I of this study. Three decisions issued within the past eighteen months give markedly different answers.

In *GEMA v. OpenAI*, discussed in Part II Section B.5 above, the Landgericht München I held in November 2025 that the training of a language model on lyrics, and the subsequent memorization of those lyrics within the model, were acts of reproduction within the meaning of the German implementation of the InfoSoc Directive, and the court rejected the application of the Article 4 CDSM exception.¹⁰⁹

On a related model-as-copy question involving a different technology, the High Court of Justice of England and Wales reached a different conclusion just a few days earlier. In *Getty Images (US) Inc. v. Stability AI Ltd.*, Mrs. Justice Joanna Smith DBE held on 4 November 2025 that the weights of Stable Diffusion, a latent diffusion model used in image generation, are not infringing copies of the photographs on which the model was trained.¹¹⁰ The plaintiff abandoned its primary infringement and database right claims at the close of trial, accepting that the evidence did not establish that the training had occurred in the United Kingdom; the dispute therefore turned on secondary infringement, which required the court to determine

jurisdictions. The figure is a moving one and is stated as of the date of this study. See also the litigation map maintained by the Copyright Alliance.

¹⁰⁹ *GEMA v. OpenAI*, LG München I, 11 November 2025 (42 O 14139/24), Endurteil. The decision is also discussed in Part II, Section B.5.

¹¹⁰ *Getty Images (US) Inc. v. Stability AI Ltd.*, [2025] EWHC 2863 (Ch) (Smith J., 4 November 2025). The decision addresses both copyright and trade mark claims; the present discussion concerns the copyright analysis only. Permission to appeal the finding on secondary infringement was granted in December 2025 ([2025] EWHC 3343 (Ch)); the appeal was pending at the date of this study.

whether the model was an article that was itself an infringing copy. The court found that the technical evidence at trial established that diffusion-model weights do not store the training data, and that, in the absence of evidence of memorization in the relevant sense, the model was not an infringing copy.¹¹¹ The court added, obiter, that had it concluded otherwise, Stability AI would have been liable on the secondary-infringement framework because the developer knew or had reason to believe that the training corpus contained works scraped from the web without consent.¹¹²

A third decision, the Hangzhou Internet Court's ruling in the so-called Hangzhou Ultraman case, addresses a related but distinct fact pattern. The defendant in that case was a platform that hosted open-source AI models to which users had access; some users had trained those models on the copyrighted Japanese manga Ultraman and uploaded the trained models for other users. The copyright owner sued the platform, and the court found the platform liable for contributory infringement. In dicta, the court expressed the view that the act of training itself might be analyzed as fair use under Chinese law, on the rationale that the purpose of using copyrighted works in the training of a generative AI model is to learn and analyze linguistic features, stylistic characteristics and other elements, rather than to reproduce the original creative expression. The court did not need to decide the point and the reasoning was not material to the outcome, but its dicta treated training as potentially eligible for fair use under Chinese law.¹¹³

The three decisions are not fully consistent on purely technical grounds because they involve different types of models (a transformer language model in Munich, a latent diffusion model in London, a platform hosting user fine-tuned image-generation models in Hangzhou) and different procedural postures. Both the Munich and the London courts wrestled with the computer-science literature on memorization in large generative models. That includes the line of work by Professors Cooper, Grimmelmann, Lemley, and others on extracting memorized content from open-weight models.¹¹⁴ That same line of work has been read by the Munich court as establishing that training amounts to large-scale reproduction, and by the London court as supporting the conclusion that diffusion-model weights are not, as a general matter, infringing copies. The comparison illustrates that the available empirical evidence on memorization does not by itself fully determine the outcome of the legal analysis. Whether the demonstrated capacity of a model to reproduce sequences from its training data should be analyzed as proof that the model is a copy of the training corpus, or as a behavioral defect to be mitigated by output filters and unlearning techniques, is a question of legal characterization on which different courts have reached different conclusions on evidentiary records that, although differing in modality and in the evidence actually before each court, drew on a similar underlying scientific literature.¹¹⁵ As noted in

¹¹¹ On diffusion-model architecture, see Part I, Section B.1, and the references there cited.

¹¹² For commentary, see E. Derclaye, "Getty Images v Stability AI: A landmark judgment reinforcing the need for the UK government to amend its copyright laws", Kluwer Copyright Blog (20 November 2025), online: <https://legalblogs.wolterskluwer.com/copyright-blog/getty-images-v-stability-ai-a-landmark-judgment-reinforcing-the-need-for-the-uk-government-to-amend-its-copyright-laws/>.

¹¹³ Hangzhou Internet Court, (2024) Zhe 0192 Min Chu No. 1587 (first instance); Hangzhou Intermediate People's Court of Zhejiang Province, (2024) Zhe 01 Min Zhong No. 10332 (second instance).

¹¹⁴ A. Feder Cooper and J. Grimmelmann, "The Files are in the Computer: Copyright, Memorization, and Generative AI", SSRN Working Paper No. 4803118 (2024). On the empirical research on memorization in language models, see also the Cooper et al. study cited at Part I, footnote 18, of this study. See also M.A. Lemley and A.F. Cooper, "Probabilistic 'Copies' in Generative AI Models", cited at Part I, footnote 25, of this study.

¹¹⁵ For a critical commentary on the GEMA reasoning at this point, see G. Frosio, "Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI", Kluwer Copyright Blog (10 and 11 December 2025), Parts 1 and 2.

Part I, whether a model contains a copy of protected works or objects of related rights should be answered model by model.

The model-as-copy question has been raised in a further and procedurally distinct form in the United States, where some plaintiffs have argued not merely that the training process produced copies but that the trained model is itself an infringing derivative work. In *Kadrey v. Meta Platforms, Inc.*, the authors contended that the LLaMA models could not function without the expressive information extracted from their books, that this information is retained within the model weights, and that the models are accordingly unauthorized derivative works, a characterization that would attach liability to the model itself without proof of a correspondence between any particular output and the works on which the model was trained. The court declined to accept that theory at the pleading stage. The defendant ultimately prevailed on fair use grounds, discussed in Section 4 below, after arguing that the weights are statistical parameters derived from the training data rather than stored expression.¹¹⁶

A comparable theory was advanced, in respect of a latent diffusion model, in *Andersen v. Stability AI Ltd.*, where the plaintiffs characterized the model as an alternative means of storing copies of the training images; that characterization has faced evidentiary difficulties in the absence of an output closely matching a protected work, a difficulty that parallels the English court's conclusion in *Getty* that the weights of a diffusion model do not store the training data. The United States experience suggests that proof that protected works were ingested in training does not, without more, establish that the trained model is a derivative of those works.¹¹⁷ As of this writing, no appellate decision had been issued on any of the above issues.

2. The text and data mining exception

The second question, which arises in the European Union and in those jurisdictions whose copyright laws have been influenced by the European framework (for example, the EEA countries), is whether the text and data mining exceptions in Articles 3 and 4 of the CDSM Directive extend to the training of generative AI models. The substantive doctrinal dispute is taken up in Part II, Section B.5, of this study. The German case law on the question is summarized here for completeness, with attention to what it teaches about the enforcement of rights holder claims.

The two German rulings in *Kneschke v. LAION e.V.* on the application of the text and data mining exception to AI training (at first instance, Landgericht Hamburg, 27 September 2024, and on appeal, Oberlandesgericht Hamburg, 10 December 2025) are discussed in Part II, Section B.5, above.

Two lessons for enforcement follow. The first is that, in jurisdictions where the TDM exception applies to AI training, the rights holder's enforcement posture is conditioned by the opt-out mechanism. Where a valid opt-out has been expressed in a manner that an automated process can detect and act upon, the rights holder may enforce her rights

¹¹⁶ For a statement of the developer position beyond the litigation context, see Google, "A Policy Agenda for Responsible Progress in Artificial Intelligence" (white paper on AI regulation), available at <https://blog.google/company-news/outreach-and-initiatives/public-policy/white-paper-ai-regulation/> accessed 7 July 2026.

¹¹⁷ On the derivative-work theory in the United States, see *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC (N.D. Cal., Chhabria J., 20 November 2023) (dismissing at the pleading stage the contention that the LLaMA models are themselves infringing derivative works, while allowing the direct infringement claim based on copying for training to proceed); and *Andersen v. Stability AI Ltd.*, No. 3:23-cv-00201-WHO (N.D. Cal., Orrick J.), orders of 30 October 2023 and 12 August 2024 (addressing, among other matters, the contention that the model stores compressed copies of the training images). The fair use disposition in *Kadrey* is discussed in Section A.4 below.

through ordinary infringement procedures. Where no such opt-out has been expressed, or where the reservation does not satisfy applicable requirements, the TDM exception may apply, subject to its other conditions, limiting an infringement claim based on the reproduction right. In that respect, the rights holder's ability to reserve rights depends partly on the technical form of the opt-out signal.

The second is that the standard of machine-readability is not static. The Hamburg Court of Appeal explicitly held that machine-readability must be assessed against the technological capabilities prevailing at the time of the disputed use. On the facts of the case, that meant the state of the art in the second half of 2021, before the widespread availability of large language models capable of extracting meaning from unstructured natural-language text. The ruling indicates that an opt-out that was not machine-readable in 2021 may be machine-readable now, and an opt-out that is not machine-readable in 2025 may become machine-readable at some future date. The practical implication for enforcement is that rights holders may need to revisit and, in many cases, restate their opt-outs as the technology evolves; the practical implication for AI developers is that the question whether the operation was lawful at the time it was carried out is not foreclosed by a subsequent change in the relevant capabilities.¹¹⁸

3. Jurisdiction over extraterritorial training

A third question, of growing practical importance, concerns the jurisdiction of national courts over training operations conducted on infrastructure located in another country. The principal Canadian ruling on the question, in *Toronto Star Newspapers Ltd. v. OpenAI Inc.*, was issued on 7 November 2025 by Kimmel J. of the Ontario Superior Court of Justice.¹¹⁹ A coalition of major Canadian media organizations had sued OpenAI, alleging that the defendant had used their content to train ChatGPT in violation of Canadian copyright law, of the technological protection measures protecting that content, and of the terms of use of the relevant websites. OpenAI moved to dismiss for lack of jurisdiction, arguing that the acts of training had been carried out on servers located in the United States and were therefore beyond the territorial scope of the Canadian Copyright Act. The court rejected that argument, holding that there was a real and substantial connection between the proceedings and Ontario and that the question whether the alleged acts engaged Canadian copyright law was a matter for the merits rather than for the jurisdictional analysis. The action proceeded to discovery in early 2026; the defendants have appealed the jurisdictional ruling, and the appeal was pending at the date of this study.

The reach of the real-and-substantial-connection test in respect of AI training is not yet established. Academic commentary on the *Toronto Star* ruling has noted that the test was developed in Canadian law in the context of the right of communication to the public, in cases such as *SOCAN v. CAIP* and *SOCAN v. ESA*, and that its transposition to the right of reproduction is not necessarily straightforward.¹²⁰ The right of reproduction has traditionally

¹¹⁸ P. Keller, "LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs", Kluwer Copyright Blog (17 and 18 December 2025), Parts 1 and 2.

¹¹⁹ *Toronto Star Newspapers Ltd. v. OpenAI Inc.*, 2025 ONSC 6217 (Kimmel J., 7 November 2025). The plaintiffs are the *Toronto Star*, the *Globe and Mail*, Postmedia, and the Canadian Broadcasting Corporation. The motion to dismiss for want of jurisdiction was based on the proposition that the alleged acts of training had been carried out on servers located in the United States and were therefore outside the territorial reach of the Canadian Copyright Act.

¹²⁰ *Society of Composers, Authors and Music Publishers of Canada v. Canadian Association of Internet Providers*, 2004 SCC 45 ("SOCAN v. CAIP"); see also *Society of Composers, Authors and Music Publishers of Canada v. Entertainment Software Association*, 2022 SCC 30. The doctrinal observation that the test was developed in respect of the right of communication to the public, and may not be readily transposable to the right of reproduction, has been advanced in academic commentary on the *Toronto Star* ruling.

been located, for jurisdictional purposes, in the place where the copy/fixation occurs. The relationship between the territorial reach of the reproduction right and the principle of national treatment under Article 5 of the Berne Convention is taken up in Section B.2 below.

Other Canadian developments illustrate the variety of procedural responses available. In *CanLII v. Caseway AI*, CanLII had sued the Vancouver-based startup Caseway AI in 2024 for the massive downloading of its database of judicial decisions for the purpose of training a legal-information AI assistant. CanLII argued that, while the underlying decisions are in the public domain, the organized structure, the metadata, and the selection of its database constitute a compilation protected by copyright. The parties settled in March 2026 on confidential terms; Because the parties settled, the case produced no judicial ruling on the merits of those issues. Class actions are also under way: in British Columbia, the author J.B. MacKinnon has filed class actions against Meta, Nvidia, Anthropic, and others, alleging that those companies have used illicit copies of works (drawn from the Books3 dataset) to train their models; in Quebec, Anne Robillard has filed for the authorization of class actions against Meta and OpenAI, seeking statutory damages of up to CAD 20,000 per work, the statutory maximum for commercial infringement. The same structural constraint proved decisive in the English Getty litigation, discussed in Section A.1 above, where the primary infringement claims were abandoned once the evidence did not establish that the training had occurred in the United Kingdom.¹²¹

In India, the High Court of Delhi upheld its jurisdiction over training conducted on servers in the United States in *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, holding, on a prima facie basis, that the storage of the works abroad was only the “terminal step in the chain of events” that begins with access from India, and that the developer’s objection was “untenable as it would lead to evasion of Indian copyright law”;¹²² the substance of the decision is discussed in Section A.4 below. In Brazil, the issue of extraterritorial training has been raised in the proceedings filed by *Folha de São Paulo* against OpenAI in August 2025, in which the newspaper alleges that the defendant has used its journalistic content in the training of ChatGPT and in the provision of services to Brazilian users without authorization.¹²³ The proceedings were under judicial secrecy; the parties settled the case in 2026, and no ruling on the merits was issued. In Argentina, *Editorial Perfil*, a major news publisher, filed suit against OpenAI and Microsoft in August 2026 before the Federal Civil and Commercial Courts, alleging the unauthorized use of its journalistic content in the training and operation of generative AI services.¹²⁴ More generally, a further argument, not yet tested before the Brazilian courts in the AI context, has been advanced by analogy to

¹²¹ *CanLII v. Caseway AI* concluded by confidential settlement in March 2026. The class actions filed by J.B. MacKinnon in British Columbia against Meta, Nvidia, Anthropic and Databricks, and by Anne Robillard in Quebec against Meta and OpenAI, are at the certification stage at the date of this study. On the statutory maximum for commercial infringement, see Copyright Act, RSC 1985, c C-42, section 38.1(1)(a).

¹²² *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, High Court of Delhi, CS(COMM) 1028/2024 (Bansal J., judgment of 24 July 2026), 2026 SCC OnLine Del 5291, at paragraph 51 for the passages quoted in the text. The ruling was delivered on the plaintiff’s application for an interim injunction; the suit remains pending, and an appeal was before a Division Bench of the High Court at the date of this study.

¹²³ *Folha de São Paulo v. OpenAI*, Case No. 1107237-96.2025.8.26.0100 (TJSP). The case was filed in August 2025 and proceeds under judicial secrecy; no public resolution had been issued at the date of this study.

¹²⁴ *Editorial Perfil S.A. v OpenAI, L.L.C., Microsoft Corporation and Microsoft de Argentina S.R.L.* (Federal Civil and Commercial Courts, Argentina, filed 24 August 2026). Little public information was available as of this writing about the complaint, but see *Editorial Perfil*, “Perfil demanda a OpenAI y Microsoft por el uso de contenidos periodísticos sin autorización” (24 August 2026) <<https://www.perfil.com/noticias/judiciales/perfil-demanda-a-openai-y-microsoft-por-el-uso-de-contenidos-periodisticos-sin-autorizacion.phtml>> accessed 29 September 2026. See also EFE, “El diario argentino Perfil demanda a OpenAI y Microsoft por uso indebido de sus contenidos” (25 August 2026) <<https://www.swissinfo.ch/spa/el-diario-argentino-perfil-demanda-a-openai-y-microsoft-por-uso-indebido-de-sus-contenidos/91953813>> accessed 29 September 2026.

Supreme Federal Court case law concerning foreign digital platforms operating in Brazil.¹²⁵ On this argument, the offering of AI services in Brazil makes the underlying training subject to Brazilian copyright law because the consequent generation of outputs by Brazilian users constitutes a continuation of the act of training; from this perspective, if the training took place unlawfully, the offering of the service that depends on it engages Brazilian copyright law through the principle of national treatment under Article 5 of the Berne Convention.¹²⁶ The argument is, in substance, a version of the principle that a violation that begins abroad but produces its effects in the forum jurisdiction may be vindicated in that jurisdiction, applied to a technological context in which the relevant fixation occurs upstream of the user-facing effect.

4. The substitution question

A fourth question concerns the legal characterization of outputs that substitute for, or compete with, the protected material on which the model was trained. The question matters because it determines whether the economic injury alleged by rights holders can be addressed within copyright's existing categories, which attach to acts of reproduction and communication to the public, or instead requires the distinct apparatus, discussed in Part II, of remuneration for substitution effects that no individual output makes actionable. The technical foundations are set out in Part I, Section C.3.

In the United States, a number of courts have already discussed the topic. In *The New York Times Company v. Microsoft Corp.*, now consolidated into *In re OpenAI, Inc. Copyright Infringement Litigation*, the defendant moved early in the proceedings to dismiss the direct copyright infringement claims on the basis that the plaintiff had used what the defendant characterized as elaborate and manipulated prompts to manufacture the regurgitation of which it complained. The District Court for the Southern District of New York declined to dismiss those claims on that ground in April 2025, and the case proceeded into discovery. The discovery phase has been dominated by a dispute over the production of consumer-facing usage logs. In a series of orders culminating in the affirmance by Stein J. on 5 January 2026 of an earlier order of Wang M.J., the court ordered the production of a randomized, de-identified sample of 20 million ChatGPT user interaction logs.¹²⁷ The order shifts the evidentiary focus from whether regurgitation can be elicited through adversarial prompting to how often it occurs in ordinary consumer use. The court treated broad, non-adversarial user-output evidence as relevant to the fair use analysis, particularly to whether the model's outputs function as market substitutes for the plaintiffs' works.

Then, in a pair of summary judgment rulings issued by the Northern District of California in June 2025, two days apart, in *Bartz v. Anthropic PBC* and *Kadrey v. Meta Platforms, Inc.*, two federal district courts reached the same outcome on the principal question, but on substantially different reasoning. In *Bartz*, Alsup J. held that the training of a large language model on lawfully acquired books was exceedingly transformative and constituted fair use under section 107 of the Copyright Act; he simultaneously held that the downloading and retention of pirated copies of those books from LibGen and PiLiMi was not fair use, and the class was certified on 17 July 2025 in respect of the piracy claims only. The parties

¹²⁵ STF, RE 1037396 and RE 1057258 (joined). The Supreme Federal Court held that platforms may be liable for content posted by users if they fail to act on notification, that platforms providing services in Brazil must comply with Brazilian legislation, and that such companies must maintain a legal representative in the country.

¹²⁶ Berne Convention for the Protection of Literary and Artistic Works, Article 5(1) (national treatment); see also Article 5(2) (independence of protection).

¹²⁷ *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143-SHS-OTW (S.D.N.Y.). The MDL consolidates a number of cases brought against OpenAI by news organizations, authors, and other plaintiffs. On the 20 million ChatGPT logs discovery order, see the references in Part I, footnote 20, of this study.

subsequently reached a US\$1.5 billion settlement, which received final approval in July 2026.¹²⁸ In *Kadrey*, Chhabria J. likewise granted summary judgment for the defendant on the fair use issue,¹²⁹ but the reasoning placed greater emphasis on the fourth fair use factor (market harm) and less on the first (transformative use), echoing the Supreme Court's methodology in *Andy Warhol Foundation v. Goldsmith*. The case has been read in subsequent commentary as turning principally on the plaintiff's failure to introduce evidence of market substitution, with the implication that a different evidentiary record could yield a different outcome.¹³⁰ The two rulings do not establish a general proposition that AI training is or is not fair use.

In *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, the United States Court of Appeals for the Third Circuit, affirming partial summary judgment on interlocutory appeal, delivered the first US appellate decision on fair use and AI training. The defendant had used the plaintiff's headnotes, editorial summaries of points of law prepared for the Westlaw research platform, to train a non-generative, AI-based legal research tool. On the first factor, the court held the use minimally transformative at best, as it served substantially the same legal-research purpose as the plaintiff's use of the headnotes; that conclusion, combined with the commercial character of the use, made the first factor weigh against fair use. It distinguished the intermediate-copying cases on the ground that the copying there was necessary to reach unprotected functional elements of software, whereas no comparable necessity justified copying the headnotes. The second factor favored the defendant but, as the court noted, rarely carries significant weight; the third weighed against fair use because entire headnotes were taken without a transformative justification. On the fourth factor, the court found harm along three dimensions: to the value of the headnotes as a draw for subscriptions to the plaintiff's platform; to the original market for legal research platforms, in which the defendant competed directly; and, at the close of its fair use analysis, to a potential derivative market for licensing headnotes as AI training data, which the court described as rapidly developing, observing that the plaintiff used its headnotes to train its own AI products, that its decision not to license others did not disprove the existence of the market, and that the unauthorized copying usurped the plaintiff's opportunity to enter it. The court distinguished the generative models at issue in *Bartz* and *In re OpenAI*, so that the transformativeness assessment in the generative cases may differ. The decision nevertheless confirms, at the appellate level, that an emerging market for licensing content that creators would generally develop or license others to develop is a cognizable consideration under the fourth factor.¹³¹ The wave of settlements in the music sector has not

¹²⁸ *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal., Alsup J., 23 June 2025) (granting in part summary judgment for the defendant on fair use grounds for training on lawfully acquired works, and denying summary judgment for the defendant in respect of the downloading and retention of pirated copies of works from LibGen and PiLiMi). The class was certified in respect of the piracy claims only. A US\$1.5 billion settlement received preliminary approval on 25 September 2025; the final fairness hearing was held on 14 May 2026 before Martínez-Olguín J., who took the matter under submission. Final approval was pending at the date of this study. The settlement received final approval from the District Court in July 2026, and the case has been closed.

¹²⁹ *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC (N.D. Cal., Chhabria J., 25 June 2025).

¹³⁰ On the contrast in the two judges' reasoning, see (among many commentaries) the analysis in the Wiggin and Dana publication of 18 September 2025, available at [wiggin.com](https://www.wiggin.com).

¹³¹ *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, No. 1:20-cv-613 (D. Del., Bibas J., 11 February 2025) (granting partial summary judgment to the plaintiff on direct infringement and rejecting the fair use defense). The district court's ruling was affirmed on interlocutory appeal: No. 25-2153 (3d Cir., argued 11 June 2026, decided 29 September 2026) (opinion by Montgomery-Reeves J.). In a footnote, the court of appeals noted the Statement of Interest filed by the United States Department of Justice in *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-3143 (S.D.N.Y., 1 September 2026), which argued that the training of a generative large language model is a transformative use, and distinguished the generative context from the case before it.

ended the litigation there. In January 2026 a group of music publishers filed a further action against the developer in the Bartz litigation, alleging the use of more than twenty thousand musical works and seeking statutory damages in excess of US\$3 billion, and collecting societies in Germany and Denmark have continued to pursue a separate music-generation developer.¹³²

Outside the United States, courts have also engaged on this topic. *Folha de São Paulo v. OpenAI*, described above, presented the question until the parties settled it. An example still pending as of this writing is the Beijing Internet Court case in which four illustrators have sued Xiaohongshu, alleging that the platform used their original works to train its AI image-generation model Trix without authorization, and that the model can generate images substantially similar to the plaintiffs' works.¹³³

In January 2025 in the Republic of Korea, the three major terrestrial broadcasters, KBS, MBC, and SBS, sued Naver before the Seoul Central District Court, alleging that their news content had been used without authorization to train Naver's HyperCLOVA family of language models. It is the first Korean proceeding in which the training question is squarely presented and, given the common ancestry of Article 35-5 of the Korean Copyright Act and section 107 of the United States Copyright Act, the United States fair use case law surveyed above is part of the arguments made before the Korean court.¹³⁴ In Brazil, a voice actor sued a shopping mall, alleging that his voice had been used without authorization in advertisements; the mall countered that the voice in question had been generated by AI, and this argument prevailed in the lower court. In 2024, however, the São Paulo Court of Appeals overturned the ruling, holding that "the use of Artificial Intelligence, in and of itself, does not eliminate (quite the contrary, it exacerbates) the risk of misuse of third-party rights", given that such tools are trained using content extracted from the web by scraping, without the authorization of the rights holders. The court also found that further evidence needed to be gathered, and the case was remanded to the lower court; it had not reached a final decision on the merits at the date of this study.¹³⁵

¹³² Universal Music Publishing Group, Concord and ABKCO filed a second lawsuit against Anthropic in the United States District Court for the Northern District of California on 28 January 2026, alleging infringement of more than 20,000 works. This action is distinct from the publishers' original 2023 lyrics case against Anthropic, which was filed in the Middle District of Tennessee and later transferred to the Northern District of California. See Music Business Worldwide, "UMG, Concord and ABKCO sue Anthropic for \$3bn in what could be single-largest non-class action copyright case" (28 January 2026) <<https://www.musicbusinessworldwide.com/umg-concord-and-abkco-sue-anthropic-for-3bn-in-what-could-be-single-largest-non-class-act/>> accessed 3 July 2026; on the 2023 case, Rolling Stone, "Music Publishers Sue Amazon-backed AI Platform for Copyright Infringement" (18 October 2023) <<https://www.rollingstone.com/music/music-news/umg-concord-abkco-sue-anthropic-ai-company-copyright-infringement-1234856761/>> accessed 3 July 2026; Music Business Worldwide, "Universal, Concord and ABKCO ask court for injunction to stop Anthropic AI from using their lyrics" (16 November 2023) <<https://www.musicbusinessworldwide.com/universal-concord-and-abkco-ask-court-for-injunction-to-stop-anthropic-ai-from-using-their-lyrics/>> accessed 3 July 2026. GEMA (Germany) and Koda (Denmark), among others, have continued to litigate against Suno notwithstanding the settlements reached by record companies with that developer and with Udio in late 2025.

¹³³ Beijing Internet Court (filed in November 2023; pending at the date of this study); the four plaintiffs are illustrators alleging that Xiaohongshu used their original works to train the AI image-generation model Trix without authorization.

¹³⁴ KBS, MBC and SBS v. Naver Corp., Seoul Central District Court (filed 13 January 2025). The claims are brought under the Copyright Act and the Unfair Competition Prevention Act, and the plaintiffs seek damages and an injunction against further use. The same broadcasters filed a comparable action against OpenAI before the same court in February 2026.

¹³⁵ Tribunal de Justiça de São Paulo, Civil Appeal No. 1119021-41.2023.8.26.0100 (6th Chamber of Private Law, Reporting Judge Costa Netto, judgment of 31 October 2024), presented at the 2025 WIPO Intellectual Property Judges Forum; see <<https://www.wipo.int/wipolex/en/judgments/details/3217>> accessed 19 September 2026.

In India, by contrast, the training question has already received the most fully reasoned judicial treatment to date outside the United States and Germany. In *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, a news agency alleged that its works had been used without authorization both to train the defendant's language models and in the generation of outputs. Ruling on the agency's application for an interim injunction, and having upheld its territorial jurisdiction over training conducted on servers located in the United States (see Section A.3 above), the High Court of Delhi held that training engages the reproduction right, since Section 14(a)(i) of the Copyright Act, 1957 extends that right to the storing of a work "in any medium by electronic means" and the deletion of copies after training does not undo the storage. The court found, however, on a *prima facie* basis, that the storage fell within the fair dealing provision of Section 52(1)(a), which covers dealings for the purposes of "private or personal use, including research". The term "research" was read in an updating manner: "the process of training LLMs underlying ChatGPT can be considered as a form of research for generating new knowledge", and "the acts of further research cannot be confined to acts of human being alone", because "ultimately the research is at the behest of humans and for the benefit of humans". Nor did the commercial character of the defendant take the use outside the provision: the court characterized the use as "purely private", the training copies remaining in a closed environment accessible to the models rather than to the public, and Section 52(1)(a) containing no exclusion of commercial uses.

On the fairness limb, the court found that the use of the works was confined to the training of the models, that the record contained no evidence of market substitution or lost revenue beyond bare averments, and that the public benefits of the technology weighed in the balance; it also noted that the plaintiff could have blocked the defendant's crawlers but had not done so. The interim injunction was refused on the further ground that any loss, if established at trial, would be compensable in damages, the court observing that a requirement to take licenses from every comparable news source would render the development of large language models economically unviable. The decision is an interim one: the suit proceeds to trial, and an appeal was pending before a Division Bench at the date of this study. Its comparative interest is nevertheless considerable. Where the German courts have asked whether a purpose-specific text and data mining exception reaches AI training, and the United States courts have asked whether training is fair use, the Delhi High Court reached the same question through a fair dealing clause enumerating "private or personal use, including research" and answered it, at this stage, in the developer's favor, on reasoning (the absence of demonstrated market harm, the availability of an unused opt-out, the public benefits of the technology) that parallels the factors weighed in the United States decisions surveyed above.

SECTION B. ENFORCEMENT AGAINST AI DEVELOPERS

The case law surveyed in Section A illustrates the substantive developments of the past eighteen months. Section B turns to the structural question of how rights holders can vindicate their claims against AI developers, given the cross-border features of the technology and the asymmetries between the parties typically involved in such disputes.

1. Structural features of the enforcement landscape

Three structural features recur. The first is the persistence and acuity of the cross-border problem identified in Section A.3 above. The training of a generative AI model is typically carried out on infrastructure located in a jurisdiction other than that of the rights holder; the model is then deployed and used in many jurisdictions simultaneously; the outputs are produced in the jurisdiction of each user. Each of those stages may engage the law of a different country. The case law to date has not yet produced a uniform answer to the question how those various national legal regimes interact, and developers may face inconsistent obligations across the jurisdictions in which their services are offered.

The second is the role of the principle of national treatment under Article 5(1) of the Berne Convention. Under the principle, authors enjoy, in countries of the Union other than the country of origin of the work, the rights that the laws of those countries grant to their own nationals, as well as the rights specially granted by the Convention. The principle is one of non-discrimination and does not of itself determine the territorial reach of national law.

The third is the use of procedural mechanisms for the aggregation of claims, of which multi-district litigation in the United States is the principal example. The consolidation of the cases brought against OpenAI by news organizations, authors, and other plaintiffs in the Southern District of New York into a single multidistrict litigation has centralized proceedings that otherwise would have been dispersed across individual actions. The Bartz class action, similarly, has aggregated claims in respect of pirated copies of approximately 482,000 books and has produced a settlement of significant scale, as noted above. In jurisdictions where class action (or “mass claim”) mechanisms or multidistrict litigation are unavailable or less developed, the cost of individual enforcement may be prohibitive, particularly for individual authors and small rights holders. The Quebec class actions described in Section A.3 are an attempt to bring the same mechanism to bear in a different procedural environment; the British Columbia class actions present a comparable picture. In jurisdictions without such mechanisms, the practical capacity of individual rights holders to enforce their rights against well-resourced developers is significantly constrained.

2. Asymmetries between the parties

Beyond these features, the case law identifies a number of asymmetries that complicate enforcement in practical terms. The first is the informational asymmetry: developers hold the technical records (dataset manifests, crawling logs, training configurations) from which the composition of the training corpus can in principle be established, even where those records are incomplete; rights holders have no comparable access, except where extraction or other empirical demonstrations produce indirect evidence. This is the subject of the transparency discussion in Section C.1 below. The second is the resource asymmetry: developers, particularly the large frontier-model developers, command resources that exceed those of most rights holders. The cost of mounting a sustained enforcement action, with the technical expert work that is now expected in such cases, is significant. The third is doctrinal uncertainty: where the law is unsettled, the published case law may be shaped by selection effects because parties settle some disputes and litigate others, while rights holders’ ability to pursue test cases is constrained by litigation costs. The resulting case law should therefore be assessed with those selection effects in mind.

SECTION C. CHALLENGES IN THE ENFORCEMENT OF RIGHTS

Three principal challenges confront rights holders seeking to enforce their rights against AI developers in the current state of the law and technology. They are addressed in turn.

1. Transparency

The first challenge is limited transparency concerning the training process. AI developers have, with very few exceptions, not disclosed the composition of their training corpora; where summary disclosures have been made, they have been at a level of generality that does not enable a rights holder to determine whether her particular work was used. This opacity is reinforced by trade secrecy claims: in early 2026, for example, the developer of a music-generation system moved to keep even the size of its training dataset under seal in litigation brought by major music companies, citing competitive harm.¹³⁶ One reason is that

¹³⁶ Music Business Worldwide, “Suno moves to keep size of its AI training data sealed in UMG and Sony’s copyright case, citing competitive harm” (2026) <<https://www.musicbusinessworldwide.com/suno-moves-to-keep->

the composition of the training corpus may be a commercial asset, and comprehensive disclosure of its contents may affect the developer's competitive position. The result is an information asymmetry that can limit the practical ability of rights holders to identify potential infringement and pursue enforcement actions.

The European Union has addressed this issue through an express statutory disclosure requirement. Article 53(1)(d) of the AI Act requires providers of general-purpose AI models to draw up and make publicly available a "sufficiently detailed summary" about the content used for the training of the model, according to a template provided by the EU AI Office.¹³⁷ Implementation turns in part on what level of detail is sufficient for the purposes of the provision. The provision is, moreover, qualified by an explicit recognition of the need to respect business secrets and confidential business information, and the implementation of the requirement is now the subject of ongoing policy and standardization work.¹³⁸ This tension between the disclosure obligation and the trade-secret carve-out is the subject of considerable debate. The eventual operational reach of the obligation is likely to depend both on the application of the template published by the AI Office in July 2025 and on the interpretation that national judicial and other authorities give to the carve-out.

Other jurisdictions have approached transparency through different instruments. In China, an extensive regime of labeling obligations applies to AI-generated content.¹³⁹ The Chinese labeling regime is not principally addressed to copyright protection but to the prevention of fraud, deception, and public confusion; nonetheless, the practical effect for rights holders has been to enable a degree of backward inference, since the labeling of AI-generated content makes it possible to identify particular outputs as AI-generated and to investigate, from the output, whether protected works are likely to have been used in training.

In the Republic of Korea, the Framework Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trust, in force since 22 January 2026, requires advance notice to users of generative AI services and the labeling of content produced by AI is required.¹⁴⁰ The labeling obligations, like those of China, are not primarily intended to protect copyright; however, they may facilitate the same kind of backward inference from outputs. Copyright issues are addressed under the Copyright Act rather than the framework legislation, an allocation between instruments that reflects the relationship, in the European Union, between the AI Act and the copyright acquis.

In Brazil, the legislative response is contained in Bill No. 2338/2023, which in its current form obliges AI developers to disclose the use of protected materials in the development of their systems through the publication of a summary; the obligation is subject to a caveat that

size-of-its-ai-training-data-sealed-in-umg-and-sonys-copyright-case-citing-competitive-harm/> accessed 7 July 2026.

¹³⁷ Regulation (EU) 2024/1689, Article 53(1)(d), referred to in Part II, footnote 82 of this study.

¹³⁸ AI Act, Recital 107, which makes plain that the obligation to publish a summary should be drafted in a manner that respects, and does not require the disclosure of, business secrets and confidential business information.

¹³⁹ Provisions on the Administration of Algorithm Recommendation Services for Internet Information Services (PRC), in force 1 March 2022; Provisions on the Management of Deep Synthesis in Internet Information Services (PRC), in force 10 January 2023; Interim Measures for the Administration of Generative Artificial Intelligence Services (PRC), in force 15 August 2023; Measures for the Labeling of Artificial Intelligence Generated and Synthesized Content (PRC), in force 1 September 2025; and the mandatory national standard of the same title approved by the State Administration for Market Regulation and the Standardization Administration in 2025.

¹⁴⁰ Framework Act on the Development of Artificial Intelligence and the Establishment of a Foundation for Trust (Republic of Korea), promulgated in January 2025 and in force since 22 January 2026.

commercial and industrial secrets must be observed.¹⁴¹ The practical effect of the disclosure obligation will depend on how broadly the trade-secret exception is interpreted. The same observation applies, *mutatis mutandis*, to the European framework.

Two observations bear on transparency as an enforcement question. The first is that transparency may facilitate enforcement but is not, by itself, sufficient. The publication of a summary, however detailed, does not on its own give rise to a cause of action; the rights holder must still establish that their work was used, that the use engaged their exclusive rights, and that no exception applies. The transparency obligation reduces the informational asymmetry but does not eliminate it. The second is that the design of the transparency obligation interacts with the design of the opt-out mechanism. A rights holder who has expressed an opt-out and whose work is identified in the transparency summary as having been included in the training corpus may have a clearer basis for an enforcement claim, subject to applicable law and the other conditions of the relevant exception; the practical question is whether the summary will be specific enough to enable that identification.

2. Data attacks, filtering, and the proliferation of opt-out standards

The second challenge concerns the techniques by which rights holders, developers, and intermediaries have sought to address the technical features of training and inference that bear on enforcement. Four distinct categories of technique can be identified.

The first is extraction attacks, developed in the security research literature and now invoked in litigation. Extraction attacks are techniques for inducing a trained model to reproduce content from its training data through carefully constructed prompting. The body of work in this area, surveyed in Part I of this study, has been cited in the security literature as evidence that extractability can be a structural property of large generative models rather than solely the result of a particular prompting strategy.¹⁴² From the perspective of rights holders, extraction attacks are an evidence-gathering technique: they may provide evidence that particular works were included in training or are recoverable from a model. From the perspective of developers, the same attacks are a vulnerability to be mitigated through deduplication of training data, output filtering, and unlearning techniques. The doctrinal question whether the manner in which a memorized output is elicited should bear on the substantive copyright analysis is now before the courts in several jurisdictions, as explained in Part I, Section C.2, of this study.

The second is the family of output and input filters. Developers have, with increasing frequency, deployed filters that prevent the model from reproducing certain identified content (output filters) or that exclude certain content from the training corpus at the data-preparation stage (input filters).¹⁴³ The role of such filters in the substantive analysis is unsettled. In the United States, the Bartz and Kadrey rulings have addressed the role of filters in different ways, with Chhabria J. in particular treating the question of market substitution as turning, in part, on the effectiveness of the filters in preventing the model from producing outputs that would substitute for specific, underlying works. In the European Union, the AI Act's transparency obligations refer to the policies that developers maintain in respect of

¹⁴¹ Bill No. 2338/2023 (Brazil), Article 62 (in the version transmitted to the Chamber of Deputies); see Part II, footnote 100 of this study for the legislative status of the bill.

¹⁴² On extraction attacks, see Part I, footnote 18 (Cooper, Gokaslan, Ahmed, Cyphert, De Sa, Lemley, Ho and Liang) and footnote 22 (Carlini et al. (2021) and Nasr et al. (2025)).

¹⁴³ Filtering at the input stage may involve excluding works whose metadata indicates a reservation of rights, deduplicating the corpus to reduce memorization risk, and excluding content from sources known to be unlicensed. Filtering at the output stage may involve detecting and blocking outputs that closely match identified protected works. The relative emphasis on input-stage and output-stage filtering varies across developers and across model types.

copyright; the implementation of those obligations may bear on the legal characterization of filtering practices as part of a developer's overall compliance posture.

The third is the family of data-poisoning techniques, by which rights holders or third parties introduce malicious, erroneous, or perturbed data into datasets in order to compromise the performance of any model trained on those datasets. In China, the Cyberspace Administration has treated data poisoning as a security threat and has launched a special action (the "Qinglang Rectification of Chaos in AI Applications") to combat it; the recommended national standard on the Basic Security Requirements for Generative Artificial Intelligence Services requires the filtering of training data to remove illegal and harmful information.¹⁴⁴ No comparable regulatory treatment of data poisoning has yet emerged in the other jurisdictions surveyed. The practical importance of data poisoning as a self-help mechanism for rights holders has not yet been authoritatively assessed in any jurisdiction; the technical and legal questions it raises are essentially open: whether the deployment of poisoning techniques against AI developers constitutes a cognizable infringement of any right held by the developer, or a defensible exercise of the rights holder's own entitlements.

The fourth is particularly relevant to European enforcement. It is the question of standards for the machine-readable expression of opt-outs.¹⁴⁵ The Hamburg Court of Appeal's ruling in *Kneschke v. LAION e.V.*, discussed in Part II, Section B.5, above, holds that a valid opt-out under Article 4(3) CDSM must be both machine-readable and machine-actionable. The technical implementation of that requirement has not yet been settled. Work on a common vocabulary for AI-related usage preferences is underway in the AI Preferences Working Group of the Internet Engineering Task Force, and was ongoing as of mid-2026; the working drafts remain works in progress, defining only a small number of use categories (in substance, AI training and search-related uses, with a category for retrieval-augmented generation under discussion), and no standard has yet been adopted.¹⁴⁶ A related question that could be worth examining is how an opt-out obligation meshes with the prohibition of formalities for the existence or exercise of rights protected under the Berne Convention. The answer may depend on the actual configuration and legal effect of the opt-out mechanism.

A number of competing standards have emerged in parallel, and they fall into two groups that are easily conflated. The first expresses only a reservation of rights, that is, an opt-out. It includes the TDM Reservation Protocol (TDMRep) and TDM-AI, both directed specifically at enabling opt-outs in compliance with Article 4(3) of the CDSM Directive; Cloudflare's content-signals mechanism, a proprietary extension of the Robots Exclusion Protocol (*robots.txt*); and broader initiatives such as C2PA and IPTC PLUS, which offer opt-out functionality through their own vocabularies. Signals of this kind are already used in practice against the crawlers of AI conversational-search engines: a substantial proportion of leading news sites are reported to block the crawler used by one such engine.¹⁴⁷ The second group

¹⁴⁴ Cyberspace Administration of China, "Qinglang · Rectification of Chaos in AI Applications" special action; *Cybersecurity Technology - Basic Security Requirements for Generative Artificial Intelligence Services*, recommended national standard, in force 1 November 2025, Article 4.2.1 (training data content filtering).

¹⁴⁵ On the LAION II ruling and the proliferation of opt-out standards, see Part II, Section B.5, and the references there to the Keller commentary.

¹⁴⁶ On the AI Preferences Working Group of the Internet Engineering Task Force, see <https://datatracker.ietf.org/wg/aipref/about/>. The TDM Reservation Protocol (TDMRep), TDM-AI, the Really Simple Licensing (RSL) Collective protocol, the Cloudflare *contentsignals.org* mechanism, C2PA, and IPTC PLUS are among the principal initiatives operating in the same space.

¹⁴⁷ A study published in 2026 found that many leading United States and United Kingdom news sites block Perplexity's crawler, PerplexityBot, in their *robots.txt* files. See BuzzStream, "Most Major News Publishers Block AI Training & Retrieval Bots" (17 December 2025, published January 2026) <<https://www.buzzstream.com/blog/publishers-block-ai-study/>> accessed 3 July 2026.

goes further, allowing a rights holder to embed affirmative, machine-readable licensing terms in the same web infrastructure. Accordingly, the question is no longer only whether a work may be used, but on what conditions, including a price and a choice between a pay-per-crawl and a pay-per-inference basis. The Really Simple Licensing (RSL) standard, an initiative supported by United States online publishers and content delivery networks, is a leading example; Cloudflare's pay-per-crawl model is another, operating at the level of the access request itself by returning a payment-required response to an unlicensed crawler. In this second group, the signal an automated agent encounters functions not only as an opt-out but as an offer to license. The introduction of AI agents could facilitate automated licensing in this space.

Compliance and adjudication are complicated by the coexistence of these two types of reservations within a single and still-unsettled standards discussion. The same provider may operate in both, as Cloudflare does through its content-signals and pay-per-crawl mechanisms, so that whether a given signal is a bare reservation or an offer to license is not always evident on its face. The European Commission has launched a consultation on protocols for reserving rights from text and data mining under the AI Act and the GPAI Code of Practice, with a view to drawing up a list of generally agreed machine-readable protocols that providers of general-purpose AI models must observe to meet their obligations under the AI Act.¹⁴⁸

The fragmentation of the standards landscape is itself a challenge to enforcement. The Hamburg Court of Appeal's ruling indicates that the machine-readability of an opt-out is to be assessed against the technological capabilities at the time of the disputed use; but in a landscape in which multiple incompatible standards coexist, the practical question of which standards a developer must observe in order to comply with the law is not straightforwardly answered. The risk identified in commentary on the ruling is that, in the absence of shared and openly governed standards, machine-readable opt-outs may become the foundation for a new intermediary layer that normalizes pay-per-crawl or pay-per-click access to the open web. Whether that risk materializes depends on choices that will be made in standardization bodies and in policy-making over the coming years.

The opt-out mechanism forms part of a rights management infrastructure with multiple layers. The first layer consists of identifiers and metadata. It contains established international standard identifiers for works and for contributors, and machine-readable rights-management and provenance metadata that can connect a work to information about its rights status. Work is underway to compile good practices for interoperable, trustworthy, and machine-readable copyright data suited to the AI environment, notably in the Copyright Infrastructure Task Force convened by the Finnish Ministry of Education and Culture¹⁴⁹, on the premise that reliable identification is a precondition both for the expression of usage preferences and for the tracing of uses after the fact. This layer is supported by the provisions of the WCT, WPPT and Beijing Treaty on rights management information mentioned above.

The second layer consist of access controls. It comprises the application programming interfaces and anti-crawling measures by which rights holders and platforms seek to ensure that only authorized agents reach their content, together with the technological protection measures used in the streaming and publishing sectors against unauthorized extraction.

¹⁴⁸ European Commission, Consultation on protocols for reserving rights from text and data mining under the AI Act and the GPAI Code of Practice (2025); see <https://digital-strategy.ec.europa.eu>.

¹⁴⁹ The Copyright Infrastructure Task Force (CITF), convened under the auspices of the Finnish Ministry of Education and Culture, published a first report, "Interoperable, trustworthy and machine-readable copyright data in the AI era", in 2025.

This layer is supported by the provisions of the WCT, WPPT and Beijing Treaty on anti-circumvention technologies mentioned above.

The third layer is attribution, by which is meant the assessment of the contribution of particular works to a given output for the purpose of allocating remuneration where applicable. Attribution in this sense should be distinguished from detection. Detection technologies seek to determine, with a degree of probability, whether a file is human- or machine-generated; they do not identify the source works and do not, on their own, support a copyright claim. Attribution technologies, by contrast, seek to estimate the share of identified works in a generated output, and for that reason depend on a degree of cooperation from, and transparency by, the developer. Investment in this capacity has begun: one provider of attribution technology, already engaged in the Swedish collective-licensing arrangement noted in Part II, was acquired by a major record company in 2026.¹⁵⁰ The attribution layer connects directly to output-level remuneration models mentioned in Part II, Section B.6. Some of those models are based not on infringement at the output stage but on the substitutive effect of outputs of certain models.

A separate family of provenance and watermarking tools, together with initiatives for the labeling of human-authored content, addresses the authenticity and origin of outputs rather than the use of protected works in training; those tools bear principally on the output questions that this study does not address, and are noted here only to situate the attribution layer within the wider infrastructure.

3. The case law as a whole

The body of case law surveyed supports several cross-cutting observations about the current enforcement of copyright in the context of generative AI training.

The first is that courts have proceeded, on the whole, with a degree of caution and have generally declined to issue rulings that purport to settle the doctrinal landscape comprehensively. The rulings to date have addressed particular technologies (latent diffusion in Getty, language models in GEMA, lawfully and unlawfully acquired books in Bartz), particular procedural postures (motion to dismiss in the New York Times litigation, summary judgment in Bartz and Kadrey), and particular questions (model-as-copy in GEMA and Getty, fair use in Bartz and Kadrey, among others). They have not yet produced a comprehensive framework in any single legal system. The German position, while now reflected in three decisions, is itself not internally uniform: the LAION line accepts the application of the TDM exception to training, while the GEMA line takes a different approach. The position in the United States is similarly stratified and awaiting appellate review, and, ultimately, one or more opinions from the United States Supreme Court.

The second is that the body of case law has nevertheless produced some points of convergence. There is convergence on the requirement, anchored in the Berne Convention, of a human author to generate the originality necessary for the protection of AI outputs. This can be observed in the United States, Kenya¹⁵¹, Germany at the trial-court level, and the European Union where the originality standard reaffirmed by the CJEU in Mio, a case

¹⁵⁰ Warner Music Group acquired the attribution technology provider Sureel in June 2026. The same provider's technology is used in the licensing arrangement involving STIM discussed in Part II, Section B.3.

¹⁵¹ Aryeh Movement Limited v. Cynthia Beldina Akoth Okello, COPTA/E001/2025 (Copyright Tribunal of Kenya) (holding that Section 22(3) of the Copyright Act (Cap 130) presupposes human involvement in the expenditure of "sufficient effort", so that work generated autonomously by an AI system cannot meet the statutory threshold unless sufficient human intervention gives the work originality).

concerning the copyright protection of furniture, presupposes a human author¹⁵². There is also convergence on the application of the reproduction right by the making of training datasets (in Germany, China, and Canada). There is, finally, some convergence on the relevance of output substitution to the analysis (in the United States and, prospectively, in the European Union through the operation of the three-step test). The points on which the case law remains divided include whether trained models contain copies of works used for training, retraining, and fine-tuning, whether and how the TDM exception extends to AI training, and how far cross-border jurisdiction reaches.

The third observation is that the practical challenges identified in Sections C.1 and C.2 above, namely transparency, the proliferation of opt-out standards, and the cost and information asymmetries between developer and rights holder, are likely to shape the development of the case law in the coming months and years. Even where the courts settle the substantive questions of copyright law, the practical enforceability of rights holder claims will depend on the mechanisms by which rights holders can ascertain the use of their works, express opt-outs, and present evidence in a form that the courts can act upon. In this domain, technical design and legal analysis are closely connected, and enforcement will be shaped by choices made in both areas.

CONCLUSION

This study identifies rapid and uneven development in two areas: jurisdictions experimenting with different combinations of exceptions, licensing mechanisms, and remuneration schemes, and a body of case law that has illuminated the issues but has not yet produced a stable framework. Member States that are now considering legislative or regulatory responses do so against a more developed landscape than was available even twelve months ago; the points on which experience and case law have converged provide a partial foundation, while the points on which they remain divided will require policy choices that this study has sought to lay out, but not to resolve. Those choices will continue to be made at the national level and, where Member States so decide, discussed at the international level through the Standing Committee on Copyright and Related Rights and the wider WIPO framework.

The study shows that the balance between the interests of rights holders and those of AI developers may be struck differently by different Member States. Five more precise questions emerge from the analysis.

The first is whether the reproductions made in assembling a corpus and training a model should require authorization, or fall within an exception or limitation, whether AI-specific, a text and data mining provision, or a general clause, and on what conditions, including lawful access, the purpose of the training, and the commercial or non-commercial character of the use.

¹⁵² *Mio and Others*, C-580/23, joined with *USM v konektra*, C-795/23 (CJEU, judgment of 4 December 2025). The joined cases concerned the copyright protection of works of applied art, specifically items of furniture (in C-580/23, the “Palais Royal” dining tables). The Court reaffirmed that subject matter qualifies as a protected work only where it is original in the sense of being the author’s own intellectual creation, reflecting free and creative choices that express the author’s personality, and that works of applied art are subject to the same originality requirement as other works. The judgment did not concern artificial intelligence; the requirement of a human author for the protection of AI outputs follows from this originality standard rather than from any ruling of the Court directed to AI. Authors may in addition face a new hurdle in enforcement proceedings, namely providing evidence that their work was not created by AI, or at least that they made enough of a contribution to justify copyright protection.

The second is the design of any reservation mechanism where they are used. This includes the form of the opt out, bearing in mind that machine-readable reservations presuppose a capacity that individual authors and performers may not possess. Conversely, a reservation regime that is not machine-actionable at scale gives developers no obvious compliance path. Compatibility of an opt-out regime with Article 5(2) of the Berne Convention should also be considered.

The third is whether compensation should attach to ingestion, to output substitution, or to both; whether it should take the form of market licensing built on exclusive rights or of a statutory remuneration claim; and, in either case, through what institutions it should flow. This includes whether collective management mandates extend to training uses, whether extended or new forms of collective licensing are needed, and how contractual chains can be designed so that value reaches the authors and performers whose works are used.

The fourth question is transparency: what record-keeping and disclosure of training material are needed to allow the exercise of rights and the administration of remuneration, and how those obligations can be reconciled with legitimate trade secrets and competitive concerns.

The fifth and final question is whether remedies remain effective and proportionate where removing works from a trained model is technically infeasible, and how jurisdiction should be exercised where training occurs abroad but the resulting services are offered domestically.

[End of document]