Rillor

EXPLAINER

What a training data license grants

Licensing data for model training, clause by clause: the licensed material, permitted uses, models and outputs, scope, term, warranties, source terms and the records each grant needs.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

What a team may do with licensed data, with the models trained on it and with their outputs depends on what the written grant names, clause by clause.

- Published license designs split "use" into separate rights. The Montreal Data License, a 2019 proposal, separates access, labelling, distribution, benchmarking, research, internal use and two kinds of commercialization, because, its authors argue, licenses tend to grant a right to use data without saying what use means [8].
- Machine-readable vocabularies separate training from retrieval. Really Simple Licensing (RSL) 1.0, a steering committee recommendation published on 10 December 2025, has one term for training or fine-tuning, one for input such as retrieval-augmented generation and grounding, and one for retrieval indexes [9].
- Rights in models and outputs are a separate question from rights in the data. The Linux Foundation's permissive Community Data License Agreement (CDLA-Permissive-2.0) places no restriction on "Results", which it defines to include machine learning models [13], while the Montreal Data License lets a licensor permit commercial use of outputs and keep the trained model itself away from third parties [8].
- End-of-term clauses can reach beyond the raw data. A March 2026 buyer-side paper from the FISD Alternative Data Council reports that some data vendors now seek deletion obligations covering "model artifacts", and proposes deleting raw data and any artifacts that could substitute for it, such as high-fidelity retrieval embeddings, without retraining models [12].
- A dataset license sits on top of the terms of its sources. A study of nearly 4,000 public datasets found that less than 33% were restrictively licensed, while over 80% of the source content in widely used text, speech and video datasets carried non-commercial restrictions [27].
- Published open licenses give the data as is. CDLA-Permissive-2.0 and Creative Commons Attribution 4.0 disclaim warranties of title and non-infringement [13, 15]. Shutterstock's 2025 annual report describes a warranty against infringement for its content licenses, with indemnification generally capped at $10,000 to $250,000, and does not say its training data licenses carry the same terms [17].
- A copyright or contract license does not settle personal data questions. Creative Commons Attribution 4.0 does not license privacy or publicity rights [15], and the European Data Protection Board said in December 2024 that models trained with personal data "cannot, in all cases, be considered anonymous" [30].
- EU law asks for authorisation where no copyright exception applies, and the U.S. Copyright Office favours a voluntary licensing market. Recital 105 of the EU AI Act says providers of general-purpose models need authorisation from rightsholders to mine works whose rights have been reserved [5], and the Office's May 2025 pre-publication report recommends letting the licensing market develop without government intervention [1].

Background

What a training data license is

In this report, a training data license is a written grant from whoever holds rights in a body of data that says what the licensee may do with it in building, testing and running models. Teams that build superintelligence (SI) systems meet two kinds in the sources reviewed. Public licenses, such as CDLA-Permissive-2.0, Creative Commons Attribution 4.0 and the Open Data Commons Attribution License, offer the same terms to anyone who follows them [13, 15, 24]. Negotiated agreements set terms between one holder and one licensee. Reddit's 2024 registration statement describes data licensing arrangements that give customers "the right to access data from our platform over the contractual period" [18], and Shutterstock's 2025 annual report describes content delivered to "our data offering partners for machine learning purposes" [17].

The U.S. Copyright Office describes voluntary licenses as agreements that "enable parties to set terms tailored to the specific uses of the works", negotiated directly or collectively [1].

Why the grant draws attention now

The Copyright Office released Part 3 of its report Copyright and Artificial Intelligence, on generative training, as a pre-publication version on 9 May 2025, and its page still listed that part as pre-publication on 7 November 2026 [2]. The report finds that licensing agreements for training "are fast emerging in certain sectors, although their availability so far is inconsistent" [1]. In the EU, the AI Act's duties for providers of "general-purpose AI models" have applied since 2 August 2025, including a copyright policy and a public summary of training content [5], and the Commission's template for that summary asks about licensed datasets in a section of its own [6].

This report is for teams that license data in for model training, for data holders who license it out and for whoever keeps the records. It is research, never legal advice: it reports what each license, standard, statute, order and filing says, with its date.

Method

What was examined. The parts of a grant for training data: the licensed material, permitted uses, rights in models and outputs, retrieval and grounding, exclusivity, territory and field of use, sublicensing and onward sharing, term and end of term, warranties and indemnities, the terms of the underlying sources, personal data and the records a grant calls for.

Sources. Public licenses and terms [13, 15, 22, 24] and their publishers' explanations [14, 16]. License designs and machine-readable vocabularies [8, 9, 10, 11]. EU and UK legislation [3, 4, 5, 29], the Commission's template and the code of practice [6, 7], the Copyright Office's report and page [1, 2] and the EDPB's opinion [30]. Three U.S. Federal Trade Commission (FTC) publications [19, 20, 21]. Two SEC filings by data holders [17, 18]. A buyer-side position paper [12]. Audits of dataset licensing [25, 26, 27, 28]. Documentation standards and platform metadata rules [23, 31, 32, 33].

How. Each source was opened on 7 November 2026 and read in the version available that day. The EU texts were read in their Official Journal versions, and later amendments were not checked. The FTC's site refused direct requests, so its three pages were opened through a web fetching tool, and every quote from them was checked word for word against Internet Archive copies of the same pages captured between 24 October and 2 November 2026. Clauses are described from what published licenses, specifications, filings and position papers say. No private agreement was available, so no statement here describes how often a clause appears in commercial licenses.

Out of scope. Prices and deal values, the comparison of named open, research-only and commercial licenses, and U.S. court decisions on fair use and training.

Findings

1. Licensing works inside copyright exceptions and opt-outs

The EU's Digital Single Market Directive of 17 April 2019 defines text and data mining as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations" [3]. Article 3 lets research organisations and cultural heritage institutions mine works they have lawful access to for scientific research. Article 4(1) allows "reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining", and Article 4(3) applies that exception only if the use "has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online" [3]. Recital 18 names metadata and the terms and conditions of a website or service as machine-readable means [3].

Article 7(1) makes "any contractual provision contrary to the exceptions provided for in Articles 3, 5 and 6" unenforceable, and Article 4 is not on that list [3]. Recital 18 adds that, in other cases, rights can be reserved "by other means, such as contractual agreements or a unilateral declaration" [3]. In the UK, section 29A of the Copyright, Designs and Patents Act 1988 permits copies for computational analysis "for the sole purpose of research for a non-commercial purpose" and makes a contract term that would prevent those copies unenforceable [4].

The EU AI Act adds duties for providers of "general-purpose AI models". Article 53(1)(c) requires a policy to comply with EU copyright law and to identify and comply with Article 4(3) reservations, and Article 53(1)(d) a public summary of the content used for training, on a template from the AI Office [5]. Recital 105 states that "Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply" [5]. Chapter V has applied from 2 August 2025, and Article 111(3) gives providers whose models were placed on the market before that date until 2 August 2027 [5].

The Commission's template, in its 5 December 2025 version, asks "Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?" and separates those datasets from other private datasets [6]. Its notice says "limited disclosure of information is required for licensed data given that the rightsholders concerned are parties to the licensing agreements" [6]. The voluntary code of practice published on 10 July 2025 says its copyright chapter "does not affect agreements between the Signatories and rightsholders authorising the use of works and other protected subject matter" [7].

In the United States, the Copyright Office's Part 3 report recommends "allowing the licensing market to continue to develop without government intervention", with targeted intervention if market failures are shown for specific types of works [1].

2. The grant clause: "use" needs a list

The Montreal Data License paper of March 2019 offers "a taxonomy for the licensing of data in the fields of artificial intelligence and machine learning" [8]. Its authors observe that licenses "posit one homogenous notion of use", and they call a bare restriction against commercial use "problematically ambiguous". They ask whether a for-profit company working for a non-profit could train a model for it and reuse that model afterward [8]. The paper defines training as exposing an untrained model to the data to adjust its weights, a trained model as one modified by that exposure and output as the results of applying a trained model to data [8]. Its rights on the data itself are access, labelling, distribution and representation. Its rights with models are benchmarking, research, publishing, internal use, output commercialization and model commercialization [8].

RSL 1.0, published by its technical steering committee as a recommendation on 10 December 2025, expresses usage terms in an XML vocabulary [9]. It defines `ai-train` as "Training or fine-tuning AI models", `ai-input` as input including "retrieval-augmented generation, grounding, or other use of the content to produce generative AI responses or search summaries", `ai-index` as inclusion in an internal index or retrieval database, and `ai-all` as any use, model evaluation included [9]. Its user classes are commercial, non-commercial, education, government and personal, and a class applies to "the entity or operational context responsible for the automated use, not the ultimate audience" [9]. Where a term both permits and prohibits the same usage, "the prohibition MUST take precedence" [9].

An IETF working document of 14 September 2026 defines "AI Training" as "Using an asset to modify the learned parameters of a generative AI model", excluding models for exclusively non-generative tasks, and "AI Use" as using an asset as input to a generative model where the user did not provide it [10]. It says that "If arrangements, such as legal agreements, exist that explicitly permit the use of that asset, those arrangements likely apply" unless they say otherwise [10]. Its note to readers says it does not reflect working group consensus [10]. A W3C community group report of 10 May 2024, which states it is not a W3C standard, expresses an Article 4 reservation with `tdm-reservation` set to 1 and points to the holder's terms with `tdm-policy` [11].

The FISD paper separates fine-tuning, which adds information to a pre-trained third-party model and "does not create a new model from scratch", from model training, which builds a new model and "represents a more intensive use of source data" [12]. It reports that data vendors "readily permit their data to be used with traditional ML models but express heightened concerns about LLM applications" [12].

Use a grant may name	Montreal Data License [8]	RSL 1.0 [9]	IETF working document [10]
Training or fine-tuning	Defined (Train), with separate rights for what follows	`ai-train`	`train-ai`, generative models only
Input at query time	No separate right among those listed	`ai-input`	`ai-use`

Use a grant may name	Montreal Data License [8]	RSL 1.0 [9]	IETF working document [10]
Retrieval index	No separate right among those listed	`ai-index`	Not defined
Evaluation	Benchmark	Within `ai-all`	Not defined
Adding labels	Labelling	Not defined	Not defined
Passing the data on	Distribute	Not defined	Not defined

Table 1. Uses named in three published vocabularies, as read for this report.

3. Models, outputs and derived data are their own grants

CDLA-Permissive-2.0 grants the right to "use, modify, and share the Data" on its terms and states that it "does not impose any restriction or obligations with respect to the use, modification, or sharing of Results" [13]. It defines Results as "any outcome obtained by computational analysis of Data, including for example machine learning models and models' insights" [13]. The Linux Foundation's June 2021 announcement says the license drops attribution as a condition of sharing [14].

The Montreal Data License builds limits into the model rights. Its research right comes "without the right to use the Output or resulting Trained Model for any purpose other than evaluating the Model Research under the same terms". Internal use lets output serve any internal purpose but not third parties. Output commercialization keeps the trained model itself from third parties, which "would allow SaaS commercialization", and model commercialization lets the model itself be made available to third parties or built into a product or service [8].

The sources reviewed do not settle whether a trained model is a derivative work of its training data. The Montreal paper calls it "highly uncertain" that models or their weights would be derivative works subject to a share-alike term [8]. Creative Commons' May 2025 guidance says that "in many cases, neither the AI model nor its outputs would be considered to be derivative works of training data under copyright law" [16]. Its guidance still takes a conservative approach, which it expects to lead to overcompliance: models and outputs shared publicly from ShareAlike data carry the same license, training copies and later use of the model stay noncommercial for NonCommercial data, and NoDerivatives works are left out of training [16]. The Copyright Office notes that training "may result in model weights that contain copies of works in the training data", in which case later copying of the weights "could also constitute prima facie infringement" [1].

The FISD paper gives FISD's Derived Data definition: data created from the licensed information that cannot be "readily reverse-engineered to recreate the Information" or used to create "a reasonable facsimile for the Information" [12]. For outputs it asks "whether the vendor's licensed data can be reverse-engineered from the output, and whether the output data could function as a material substitute" [12].

4. Retrieval and grounding copy the work again

The Copyright Office states that retrieval-augmented generation "also involves the reproduction of copyrighted works", through material copied into a retrieval database or retrieved at query time [1]. The vocabularies keep it apart from training: RSL's `ai-input` and `ai-index` are separate terms from `ai-train` [9], and the IETF document's use category is separate from its training category [10]. At the end of a license, the FISD paper treats "high-fidelity embeddings designed for Retrieval-Augmented Generation" as artifacts that could serve as functional substitutes for the raw data [12]. Reddit's filing lists licensing purposes "including machine learning, business analysis, display and training generative AI models" [18].

5. Scope: exclusivity, territory, field of use and onward sharing

Exclusivity. Shutterstock's annual report says content from its library is "typically" licensed non-exclusively and custom content exclusively [17]. Creative Commons Attribution 4.0 and Common Crawl's terms both grant non-exclusive licenses [15, 22].

Territory. Creative Commons Attribution 4.0 grants a worldwide license [15]. RSL has a `geo` type that "Restricts usage to within specified ISO 3166-1 alpha-2 country or region codes" [9]. The AI Act's recital 106 ties the provider's copyright policy to the model regardless of where training takes place [5].

Field of use. The Montreal paper lists added restrictions a licensor may impose: designated third parties, attribution or confidentiality, and ethical limits such as "exclusions related to use in health-related fields, or restrictions against military use" [8]. RSL's user classes divide by who runs the automated use [9].

Sublicensing and onward sharing. The Montreal paper notes that restricting sublicensing "would also prevent third party contractors from working with the licensed data" [8]. Common Crawl grants "a limited, non-assignable, non-transferable, non-sublicensable, non-exclusive, limited license" [22], and Creative Commons Attribution 4.0 is non-sublicensable [15]. CDLA-Permissive-2.0 lets a recipient share the data, modified or not, if the agreement text goes with it [13]. Reddit's filing says its arrangements "include protections against abuse and misuse of Reddit data" [18]. The IETF document lists as an open issue whether compliance includes passing on a dataset's preferences when a model trained on it is distributed [10].

6. Term, termination and what happens to copies and models

Reddit's filing describes data licensing arrangements entered into in January 2024 with terms of two to three years, a right of access over the contractual period, a fixed or usage-based fee, delivery "through continuous access to our data API as well as quarterly transfers of Reddit data", and API performance requirements [18].

Some licenses run for as long as the rights last. Shutterstock offers the content on its websites under perpetual licenses, apart from some custom, editorial and music content with special terms, and licensed content stays "perpetually subject to the customer's license even if the contributor removes the image" from the marketplace, apart from periodic removals over third-party rights concerns [17]. Creative Commons Attribution 4.0 applies for the term of the licensed rights, ends automatically on a breach and is reinstated if the breach is cured within 30 days of its discovery [15].

The FISD paper describes expiry from the buyer side. Vendors "have traditionally required customers to delete raw licensed data upon agreement expiration", and "some now seek broader obligations that extend to 'model artifacts' which may embed licensed data" [12]. Vendors may be concerned that content "becomes permanently embedded in the model parameters, making post-termination data deletion ineffective" [12]. The paper proposes that firms certify deletion of raw datasets and remove functional substitutes such as retrieval embeddings, and calls retraining models to excise historical use "commercially impracticable" [12]. The paper is one side's position. In the Montreal design, a licensee holding only an evaluation right could test models on the data but could not reuse a model as modified by it, so "the weights of the Trained Model would have to be deleted" [8].

The FTC has ordered models deleted in settlements over data it said was collected or used deceptively or illegally [19, 20]. The FTC's final order against a photo app developer, announced on 7 May 2021, required it to delete photos and videos of users who deactivated their accounts and "the models and algorithms it developed by using the photos and videos uploaded by its users" [19]. Its March 2022 settlement with WW International and its Kurbo app required deletion of personal information illegally collected from children under 13, destruction of "any algorithms derived from the data" and a $1.5 million penalty [20]. Both orders concern data the FTC said was collected or used deceptively or illegally. Neither concerns the expiry of a license.

7. Warranties and indemnities

The public licenses reviewed give the data as is. CDLA-Permissive-2.0 disclaims warranties of "TITLE, NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE" [13], and Creative Commons Attribution 4.0 offers the material "as-is and as-available" with the same exclusions [15]. Creative Commons Attribution 4.0 also states that it does not license moral rights, publicity, privacy or similar personality rights, or patent and trademark rights [15].

Shutterstock's annual report describes another arrangement. For its content licenses, it represents and warrants that unaltered content used in compliance with its license agreements and applicable law does not infringe copyright, trademark or other intellectual property rights or violate privacy or publicity rights, and its agreements "generally cap our indemnification obligations at amounts ranging from $10,000 to $250,000", with some uncapped products [17]. The filing reports increased demand for its metadata for model training [17]. It does not say its training data licenses carry the warranty or the caps.

The Montreal paper adds a limit no warranty clause removes: data "may be subject to other rights such as privacy rights, image rights and publicity rights", and a dataset of ill-acquired data is not saved by the license placed on it [8].

8. Layered terms: the dataset license and the terms of its sources

Common Crawl's terms, last updated on 7 March 2024, license access to its service and state that the crawled content "may be subject to separate terms of use or terms of service from the owners of such Crawled Content" [22]. Common Crawl may amend its terms at any time [22]. FineWeb, a web text dataset, is released under the Open Data Commons Attribution License v1.0, and its card adds that "The use of this dataset is also subject to CommonCrawl's Terms of Use" [23]. That license, in turn, "only governs the rights over the Database, and not the contents of the Database individually" [24].

The Commission's template makes the same point for free datasets: public availability for free "does not mean that the content at issue is necessarily free of rights", and certain open licenses may prohibit uses relating to model training [6].

Audits measure the gap. A 2024 study of nearly 4,000 public datasets from 1990 to 2024 found that while less than 33% were restrictively licensed, over 80% of the source content in widely used text, speech and video datasets carried non-commercial restrictions [27]. A 2023 audit of more than 1,800 text datasets found that three widely used hosting sites labelled license use cases too permissively in 29%, 27% and 16% of cases, which its authors attribute to contributors mistaking code licenses for data licenses [25]. A 2024 audit of 14,000 web domains found terms of service crawling restrictions on 45% of the C4 corpus and inconsistencies between sites' terms of service and their robots.txt files [26].

The Common Pile v0.1 team, which assembled eight terabytes of openly licensed text from 30 sources in 2025, names the failure "license laundering", in which a work is redistributed with an incorrect license, typically by someone who does not hold the rights [28]. They kept only sources where they were confident the license came from the copyright holder, and left out NonCommercial and NoDerivatives content, which the Open Definition 2.1 does not count as open [28]. The team also notes that an open license does not necessarily mean the rights holder contemplated training [28].

9. Personal data sits outside a copyright or contract grant

The GDPR requires that personal data be "processed lawfully, fairly and in a transparent manner" and "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" [29]. The Open Data Commons Attribution License names privacy and data protection rights among those it does not address [24], and Creative Commons Attribution 4.0 does not license privacy rights [15].

The EDPB's Opinion 28/2024, adopted on 17 December 2024, states that "AI models trained with personal data cannot, in all cases, be considered anonymous" [30]. Where a model that retains personal data is deployed by another controller, the opinion says supervisory authorities should consider whether the

deployer assessed that the model "was not developed by unlawfully processing personal data" [30].

A February 2024 post on the FTC's Office of Technology blog warns that adopting more permissive data practices, such as using data for training, through a "surreptitious, retroactive amendment" to terms of service or a privacy policy may be unfair or deceptive [21]. The post, a staff blog post by the Office of Technology and the Division of Privacy and Identity Protection and not a rule, says a business "cannot then unilaterally renege on those commitments after collecting users' data" [21].

10. The record each grant needs

Datasheets for Datasets, published in Communications of the ACM in December 2021, asks whether a dataset is distributed under a copyright or other intellectual property license or terms of use, asks for a link to the terms and any fees, and asks "Have any third parties imposed IP-based or other restrictions on the data associated with the instances?" [31]. MLCommons' Croissant format, version 1.0 of 1 March 2024, gives a dataset a `license` property that may hold several values and recommends the URL of a known license from the SPDX list, alongside a version number and file checksums [32]. One dataset hub reads a dataset's license from its card's metadata, offers identifiers that include "unknown", and asks for the license text in a LICENSE file when the identifier is "other" [33].

Some terms set what travels with the data. CDLA-Permissive-2.0 requires that the agreement text be made available with shared data [13]. RSL's attribution type requires "visible credit and a functional link to the original source" [9], and Creative Commons' guidance says attribution for training "could be a simple link to the source of the dataset used to train the model" [16].

The 2023 audit found that aggregators left 69% to 72% of licenses unspecified, against 30% under the audit's own protocol [25].

Implications

For teams licensing data in. A grant that lists each intended use in its own words leaves less to interpretation than a grant of "use" [8]. The three vocabularies in Table 1 work as a checklist: training, fine-tuning, input at query time, retrieval indexes, evaluation, labelling and onward sharing [8, 9, 10]. Published designs treat rights in models and in outputs apart from rights in the data [8, 13]. The end-of-term clause decides what happens to raw copies, embeddings and model artifacts, and the published positions on it differ [8, 12]. The terms beneath a dataset license come with the data [22, 23, 27].

For data holders licensing out. The same vocabularies let a holder state which uses it grants and which it reserves, and RSL resolves a conflict in favour of the prohibition [9]. In the EU, a reservation over content made publicly available online is expected in machine-readable form [3]. Commitments a holder has made to the people in its data limit what it can grant [21, 29].

For whoever keeps the records. A license kept per record answers the questions that datasheets, the Commission's template and the audits ask [6, 25, 31]. A Croissant `license` field with an SPDX URL, a version and checksums ties a dataset to the terms it came under [32]. A copyright or contract license does not license privacy rights [15], and the GDPR ties personal data to the purposes it was collected for [29].

Limits of this analysis

- No private training data agreement was reviewed. The report states nothing about how often exclusivity, deletion, output, audit or indemnity clauses appear in commercial licenses.
- The machine-readable vocabularies are not adopted standards. RSL 1.0 is a steering committee recommendation [9], the IETF document is a working document without consensus [10] and TDMRep is a community group report [11]. No adoption figures were found for them, for the Montreal Data License or for CDLA-Permissive-2.0.
- The Copyright Office's Part 3 is a pre-publication version [2].
- Shutterstock's perpetual, non-exclusive, warranty and indemnity terms describe its content licenses, and the filing does not say its data deliveries for machine learning carry them [17]. The FISD paper is one buyer group's position [12].
- The sources disagree on whether trained models are derivative works of their training data, and the report draws no conclusion [1, 8, 16].
- Jurisdictions other than the EU, the UK and the U.S. federal level were not reviewed, nor were U.S. state laws or court decisions.

Sources

1. Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version). U.S. Copyright Office, May 2025, released 9 May 2025, viewed 7 November 2026. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

2. Copyright and Artificial Intelligence. U.S. Copyright Office, web page, viewed 7 November 2026. https://www.copyright.gov/ai/

3. Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market. European Parliament and Council, 17 April 2019, Official Journal of the European Union L 130, 17 May 2019, viewed 7 November 2026. https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng

4. Copyright, Designs and Patents Act 1988, section 29A: Copies for text and data analysis for non-commercial research. UK Government, legislation.gov.uk, inserted 1 June 2014 by S.I. 2014/1372, viewed 7 November 2026. https://www.legislation.gov.uk/ukpga/1988/48/section/29A

5. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Parliament and Council, 13 June 2024, Official Journal of the European Union, 12 July 2024, viewed 7 November 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

6. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, AI Office, page published 24 July 2025, Communication C(2025) 8311 final, 5 December 2025, viewed 7 November 2026. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models (PDF: https://ec.europa.eu/newsroom/dae/redirection/document/118480)

7. Code of Practice for General-Purpose AI Models, Copyright Chapter. European Commission, independent chairs of Working Group 1, published 10 July 2025, viewed 7 November 2026. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (PDF: https://ec.europa.eu/newsroom/dae/redirection/document/118115)

8. Towards Standardization of Data Licenses: The Montreal Data License. Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio and Alex Shee, arXiv 1903.12262, 21 March 2019, viewed 7 November 2026. https://arxiv.org/abs/1903.12262

9. Really Simple Licensing (RSL) 1.0 Specification, RSL-SPEC-1.0. RSL Technical Steering Committee, 10 December 2025, viewed 7 November 2026. https://rslstandard.org/rsl

10. A Vocabulary For Expressing AI Usage Preferences, draft-ietf-aipref-vocab-08. P. Keller and M. Thomson (ed.), IETF AI Preferences Working Group, Internet-Draft, 14 September 2026, viewed 7 November 2026. https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/

11. TDM Reservation Protocol (TDMRep), Final Community Group Report. W3C Text and Data Mining Reservation Protocol Community Group, 10 May 2024, viewed 7 November 2026. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

12. FISD Alternative Data Council Buy-Side Perspective: Understanding Investment Firms' Use of Licensed Data with GenAI, Version 1.0. FISD Alternative Data Council, Software & Information Industry Association, March 2026, viewed 7 November 2026. https://fisd.net/wp-content/uploads/2026/03/FISD-Alternative-Data-Council-Buy-Side-Perspective-Understanding-Investment-Firms-Use-of-Licensed-Data-with-GenAI-1.pdf

13. Community Data License Agreement, Permissive, Version 2.0. The Linux Foundation, 2021, viewed 7 November 2026. https://cdla.dev/permissive-2-0/

14. Enabling Easier Collaboration on Open Data for AI and ML with CDLA-Permissive-2.0. The Linux Foundation, press release, 22 June 2021, viewed 7 November 2026. https://www.linuxfoundation.org/press/press-release/enabling-easier-collaboration-on-open-data-for-ai-and-ml-with-cdla-permissive-2-0

15. Creative Commons Attribution 4.0 International Public License, legal code. Creative Commons, viewed 7 November 2026. https://creativecommons.org/licenses/by/4.0/legalcode.en

16. Using CC-licensed Works for AI Training. Creative Commons, May 2025, viewed 7 November 2026. https://creativecommons.org/using-cc-licensed-works-for-ai-training/

17. Shutterstock, Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, EDGAR, filed 17 February 2026, viewed 7 November 2026. https://www.sec.gov/Archives/edgar/data/1549346/000154934626000008/sstk-20251231.htm

18. Reddit, Inc. Registration Statement on Form S-1. U.S. Securities and Exchange Commission, EDGAR, filed 22 February 2024, viewed 7 November 2026. https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm

19. FTC Finalizes Settlement with Photo App Developer Related to Misuse of Facial Recognition Technology. U.S. Federal Trade Commission, press release, 7 May 2021, viewed 7 November 2026. https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology

20. FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids' Sensitive Health Data. U.S. Federal Trade Commission, press release, 4 March 2022, viewed 7 November 2026. https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive

21. AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive. U.S. Federal Trade Commission, Office of Technology blog, 13 February 2024, viewed 7 November 2026. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive

22. Terms of Use. Common Crawl Foundation, last updated 7 March 2024, viewed 7 November 2026. https://commoncrawl.org/terms-of-use

23. FineWeb dataset card. Hugging Face, HuggingFaceFW/fineweb, viewed 7 November 2026. https://huggingface.co/datasets/HuggingFaceFW/fineweb

24. Open Data Commons Attribution License (ODC-By) v1.0. Open Data Commons, Open Knowledge Foundation, viewed 7 November 2026. https://opendatacommons.org/licenses/by/1-0/

25. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre, Robert Mahari, Anthony Chen and others, arXiv 2310.16787, version 3, 4 November 2023, viewed 7 November 2026. https://arxiv.org/abs/2310.16787

26. Consent in Crisis: The Rapid Decline of the AI Data Commons. Shayne Longpre, Robert Mahari, Ariel Lee and others, arXiv 2407.14933, version 2, 24 July 2024, viewed 7 November 2026. https://arxiv.org/abs/2407.14933

27. Bridging the Data Provenance Gap Across Text, Speech and Video. Shayne Longpre and others, arXiv 2412.17847, ICLR 2025, version 2, 19 February 2025, viewed 7 November 2026. https://arxiv.org/abs/2412.17847

28. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Nikhil Kandpal and others, arXiv 2506.05209, 5 June 2025, viewed 7 November 2026. https://arxiv.org/abs/2506.05209

29. Regulation (EU) 2016/679 (General Data Protection Regulation). European Parliament and Council, Official Journal of the European Union L 119, 4 May 2016, viewed 7 November 2026. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

30. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. European Data Protection Board, adopted 17 December 2024, viewed 7 November 2026. https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en (PDF: https://www.edpb.europa.eu/system/files/documents/2024-12/edpb_opinion_202428_ai-models_en.pdf)

31. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford, arXiv 1803.09010, version 8, 1 December 2021, Communications of the ACM, December 2021, viewed 7 November 2026. https://arxiv.org/abs/1803.09010

32. Croissant Format Specification, Version 1.0. MLCommons, 1 March 2024, viewed 7 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-spec.html

33. Licenses. Hugging Face Hub documentation, viewed 7 November 2026. https://huggingface.co/docs/hub/repositories-licenses

About this report

Published by Rillor on 8 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. What a training data license grants. 8 November 2026. https://rillor.com/insights/what-a-training-data-license-grants

Work with Rillor

Send the data you need, the publishers or holders you know of, and the use the license must cover. Contact us at rillor.com/contact?type=acquisition or contact@rillor.com. More on acquisition and licensing: rillor.com/data#acquisition