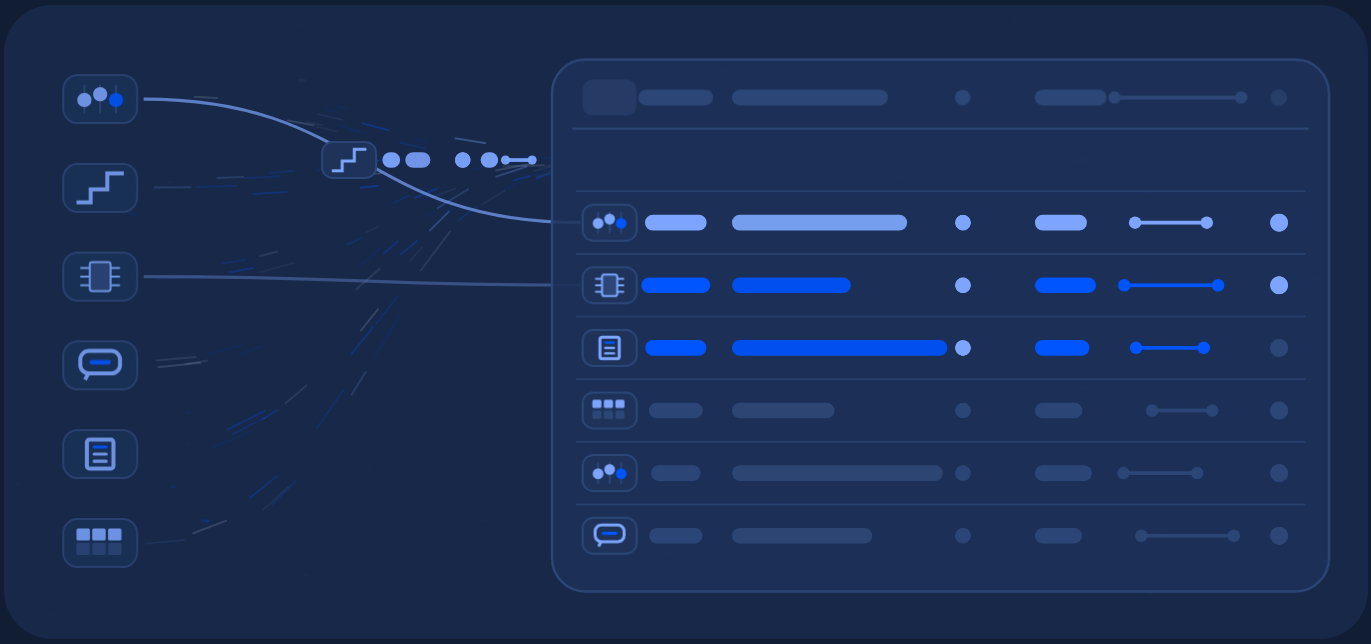
Rillor

METHODS NOTE

Buying training data: what to check before you license a dataset

Before you buy training data, check its datasheet, the rights back to each source, what the license allows for models, the personal data inside and a measured sample.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A buyer of training data can check five things before signing and again after each delivery: the documentation, the rights back to each source, what the license allows for models, the personal data inside and a measured sample.

- License labels on dataset platforms are weak evidence. A 2024 audit of 1,858 fine-tuning datasets found more than 70% of licenses for popular datasets on GitHub and Hugging Face "unspecified", and platform labels too permissive in 16% to 29% of cases [6].
- A dataset's own license can understate the restrictions on its sources. Across nearly 4,000 public text, speech and video datasets, under 33% were restrictively licensed, while over 80% of their source content carried non-commercial restrictions [7].
- A datasheet, the dataset documentation proposed in Datasheets for Datasets, answers when the data was collected, whether the people in it consented, which third-party restrictions apply and how versions change [1, 3, 4].
- Licenses written for machine learning separate research use, internal use of outputs, selling outputs and shipping the trained model, and some data vendors now seek deletion, when a license ends, of "model artifacts" which may embed licensed data [9, 12].
- How the data was obtained matters apart from how it is used. The U.S. Copyright Office's pre-publication report says knowing use of pirated works should weigh against fair use, and a June 2025 district court order found pirated library copies not justified by fair use and set them for trial [13, 14].
- Personal data carries the promises it was collected under. FTC orders have required deletion of models and products built from improperly used data, and one study estimated that 15 demographic attributes re-identify 99.98% of Americans in any dataset [18, 19, 22].
- A clean sample has a measurable limit. Zero defects in 200 records still allows a true defect rate up to 1.83% at 95% confidence, derived from NIST's exact binomial method [31].
- Diligence goes stale. SEC examination staff found investment advisers that onboarded alternative data providers with no system for deciding when to repeat diligence as time passed or collection practices changed [34].

Background

Buying training data here means licensing a dataset that someone else assembled, under a written agreement, for model training. The buyer inherits what the seller did upstream. NIST's profile for generative systems lists procured datasets among third-party components that "might be improperly obtained or not properly vetted, leading to diminished transparency or accountability for downstream users" [5].

Documentation drifts as data moves. The 2024 license audit found collections "treated as monoliths, rather than a lineage of data sources", often repackaged and relicensed by successive practitioners [6]. Each round can drop or change a restriction.

Licensing for model training is a live practice. The U.S. Copyright Office wrote in May 2025 that "Voluntary licensing is already happening in some sectors, and it appears reasonable or likely to be developed in others" [13]. In its view, where licensing options exist or are likely to be feasible, that consideration disfavors fair use under the fourth factor [13].

This note checks the dataset and its license from the buyer's side, before signing and after each delivery.

Method

What was examined. Documentation standards for datasets, data licenses and the rights they grant for models and outputs, regulator and court records on how data was obtained and on personal data, audits of public datasets, and the disclosure duties two laws place on developers.

Sources. Thirty-six sources, each opened on 7 November 2026 and read for the passages cited: 9 peer-reviewed papers, 3 preprints, 5 standards and specifications, 2 data licenses or license guides, 2 industry papers, 5 EU and California legal texts and guidance documents, 9 regulator and court records, and one marketplace's documentation.

How. Each check comes from the source that states it, put as a buyer's question. Figures marked "derived" are arithmetic on a source's formula [31]. No dataset, seller or model is ranked, and model developers named in sources are described generically.

Weight. Staff publications are read as staff views [20, 34], the Copyright Office report as its pre-publication view [13] and the court order as one district court's ruling on summary judgment [14]. The two papers from FISD, part of the Software & Information Industry Association, address investment firms [12, 35]. The opinion of the European Data Protection Board (EDPB) concerns models and is used by analogy for the data behind them [21].

Findings

1. What a listing and a datasheet should state

Datasheets for Datasets asks that each dataset "be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on" [1]. Its seven sections follow the dataset lifecycle from motivation to maintenance, and four matter to a buyer in particular [1]:

- Composition: "Are there any errors, sources of noise, or redundancies in the dataset?", and whether external resources the dataset relies on carry "restrictions (e.g., licenses, fees)".
- Collection: the timeframe of collection, and whether the individuals were notified and consented.
- Distribution: the license or terms of use, "IP-based or other restrictions" imposed by third parties, and export controls.
- Maintenance: errata, updates that correct labels or delete instances, whether older versions stay hosted, and retention limits on data about people.

Croissant, MLCommons' metadata format (version 1.0, March 2024), makes parts of this machine-readable [3]. Its license field recommends the URL of a known license, such as one on the SPDX list [3]. It recommends MAJOR.MINOR.PATCH versions, where "If the MAJOR version is incremented, the existing data has been changed (edited, removed or shuffled records across splits), or extended in a way which doesn't allow for easy access to the data as it was at the previous version" [3]. Each file can carry a sha256 checksum, and data that is not static sets isLiveDataset to true [3]. Croissant RAI adds responsible-use fields such as rai:personalSensitiveInformation, rai:dataUseCases, rai:dataCollectionTimeFrame and rai:annotationProtocol [2].

The Data & Trust Alliance's Data Provenance Standards (v1.0.0, July 2024) "consist of 22 metadata fields, grouped into 3 standards: Source, Provenance, and Use" [4]. They include data origin geography, the range of dates for data generation, consent documentation location, license to use and intended data use [4]. The Alliance says the standards "do not guarantee compliance with any particular data protection law" [4].

One cloud marketplace's guidelines say descriptions "must be accurate, contain valid contact information, and note if any data has been aggregated or anonymized", and they put use terms in the subscription agreement [25]. NIST's profile suggests that developers "Document training data curation policies, to the extent possible and according to applicable laws and policies" [5], one more document a buyer can request.

Buyer question	Datasheets section [1]	Croissant field [2, 3]	Data Provenance Standards field [4]
Which license applies?	Distribution	license	License to use

Buyer question	Datasheets section [1]	Croissant field [2, 3]	Data Provenance Standards field [4]
When was it collected?	Collection	rai:dataCollectionTimeFrame	Range of dates for data generation
Where did it come from?	Collection	rai:dataCollection	Data origin geography
What personal data is in it?	Composition, Collection	rai:personalSensitiveInformation	Consent documentation location
What is it meant for?	Uses	rai:dataUseCases	Intended data use
How does it change?	Maintenance	version, isLiveDataset	Date of previously issued version of the dataset

2. The rights chain back to the source

The 2024 audit traced "1,858 finetuning datasets from 44 of the most widely used and adopted text data collections" [6]. On GitHub, Hugging Face and Papers with Code, 72%, 69% and 70% of licenses were unspecified, against 30% under the audit's own protocol [6]. "66% of analysed Hugging Face licences were in a different use category, often labelled as more permissive than the author's original licence" [6]. The platforms labeled license use too permissively in 29%, 27% and 16% of cases, often because contributors mistook a repository's code license for a data license [6]. The audit also found that "85% of dataset licences request attribution, and 30% include a share-alike clause" [6], conditions that travel with the data.

A 2025 study traced "nearly 4000 public datasets between 1990-2024, spanning 608 languages, 798 sources, 659 organizations, and 67 countries" [7]. Under 33% were restrictively licensed, while over 80% of their source content carried non-commercial restrictions [7]. That gap is the reason to ask for the sources as well as the dataset's license.

FISD's identification factors list "requesting supplemental documentation from the vendor (e.g., data dictionary, upstream agreements, compliance policies)" among the ways to verify a vendor [35]. NIST suggests categorizing content by its third-party rights, such as copyright, intellectual property and data privacy [5]. Open licenses give no promise about the chain: the Community Data License Agreement CDLA-Permissive-2.0 provides data "AS IS" and disclaims "ANY WARRANTIES OR CONDITIONS OF TITLE, NON-INFRINGEMENT" [10]. Whether a paid license warrants title is a question for each agreement.

3. How the data was obtained: copyright and crawl rules

The Copyright Office's pre-publication report of May 2025 treats acquisition as its own question: "the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use without being determinative" [13]. It adds that "Copyright owners have a right to control access to their works, even if someone seeks to obtain them in order to make a fair use" [13].

A June 2025 order of the U.S. District Court for the Northern District of California, in a suit by authors against a model developer, granted summary judgment "that the training use was a fair use" and held the print-to-digital format change of purchased books a fair use as well [14]. On pirated copies it denied summary judgment and set a trial, writing: "The downloaded pirated copies used to build a central library were not justified by a fair use. Every factor points against fair use" [14]. It is one district court's ruling [14].

Crawl permissions carry a date. An audit of 14,000 web domains found that in 2023 to 2024 new restrictions left "~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use" [8]. C4 is the Colossal Clean Crawled Corpus of web text [27]. "For Terms of Service crawling restrictions, a full 45% of C4 is now restricted", in data collected through April 2024, and it found general inconsistencies between sites' terms of service and their robots.txt [8].

The EU's General-Purpose AI Code of Practice is "a voluntary tool, prepared by independent experts in a multi-stakeholder process" [36]. Its copyright chapter commits signatories to respect "any technological denial or restriction of access imposed by subscription models or paywalls", to exclude sites that courts or public authorities in the EU and the EEA recognize as "persistently and repeatedly infringing copyright and related rights on a commercial scale", and to use crawlers that follow robots.txt as specified in RFC 9309 [17]. The commitment "does not affect the application of Union law on copyright and related rights to protected content scraped or crawled from the internet by third parties" [17]. FISD adds that a dataset "may have heightened risk implications if the data is gathered using processes (e.g., web scraping)" [35]. A buyer of crawled data can put the code's three crawl commitments, as questions, to the seller.

4. What the license grants for models and outputs

The Montreal Data License paper "provides a taxonomy for the licensing of data in the fields of artificial intelligence and machine learning" [9]. It names ten rights, including Research, Internal Use, Output Commercialization and Model Commercialization, and argues that "research" is ambiguous in license language [9].

Right [9]	What it allows	Question for the buyer
Research	Use the data to create or improve models, with output and trained model used only to evaluate the research	Is training allowed beyond evaluation?

Right [9]	What it allows	Question for the buyer
Internal Use	Use the output internally for any purpose, never for third parties	Can outputs feed internal work?
Output Commercialization	Make the output available to third parties, with the trained model kept back	May outputs reach clients?
Model Commercialization	Make the trained model available to a third party or embody it in a product or service	May the model ship?

CDLA-Permissive-2.0 "does not impose any restriction or obligations with respect to the use, modification, or sharing of Results", defined as "any outcome obtained by computational analysis of Data, including for example machine learning models and models' insights" [10]. Creative Commons states that "CC licenses are copyright licenses" that "apply only when copyright permission is required" [11]. Its guidance reads NonCommercial as covering "all stages, from copying the data during training to sharing the trained model", NoDerivatives as keeping content out of training data, and ShareAlike as requiring the same license on publicly shared models or outputs based on such content [11]. CC calls that guidance "likely to lead to overcompliance" [11].

FISD's March 2026 buy-side paper reports "the conflation of different AI technologies, particularly the distinction between traditional machine learning (ML) and LLMs" in license negotiations [12]. It defines fine-tuning as adapting an existing model and model training as building a new one, "a more intensive use of source data" [12]. On termination it reports: "While vendors have traditionally required customers to delete raw licensed data upon agreement expiration, some now seek broader obligations that extend to 'model artifacts' which may embed licensed data" [12]. Its own position is that firms can certify deletion of raw datasets and remove artifacts "that could serve as functional substitutes for the raw dataset, such as high-fidelity embeddings designed for Retrieval-Augmented Generation" [12].

Model-generated parts bring other terms. About 12% of the datasets in the 2024 audit were annotated with one commercial model provider's service, whose terms bar using outputs to build competing models [6]. The authors note such terms "constitute a contractual agreement, not a copyright licence, potentially making them unenforceable against third parties who did not create the data" [6]. They also found that "a full 45% of NC/A-O datasets are synthetic, compared to <14% in more permissive licence categories", where NC/A-O marks non-commercial or academic-only licenses [6]. NIST suggests contracts that "specify content ownership, usage rights, quality standards, security requirements, and content provenance expectations" [5].

5. Personal data and the promises it was collected under

Datasheets ask whether people were notified and consented [1], Croissant RAI describes personal and sensitive information [2], and the Data Provenance Standards record where consent documentation sits [4].

FTC orders show the downstream cost. In 2021 the final order with a photo app developer required it to delete "the models and algorithms it developed by using the photos and videos uploaded by its users" [19]. In 2024 the final order with a seller of location data required a supplier assessment program to make sure its suppliers obtained informed consent, and deletion of earlier data and "any products developed from this data" unless consent was obtained or the data was de-identified [18].

FTC staff wrote in February 2024 that "It may be unfair or deceptive for a company to adopt more permissive data practices", such as sharing data with third parties or using it to train models, "and to only inform consumers of this change through a surreptitious, retroactive amendment to its terms of service or privacy policy", and that a company "cannot then unilaterally renege on those commitments after collecting users' data" [20]. The post is a staff blog [20].

The EDPB's Opinion 28/2024 says supervisory authorities should consider "whether the controller deploying the model conducted an appropriate assessment" that the model "was not developed by unlawfully processing personal data" [21]. The controller "should consider, for instance, if the data originates from a personal data breach or if the processing was subject to the finding of an infringement from a SA or a court" [21]. The opinion concerns models. For high-risk systems, the EU AI Act's data governance rules cover "data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection" [15].

De-identified data needs its own scrutiny. A 2019 study estimated that "99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes" [22]. Extraction attacks on a public language model from 2019 recovered names, phone numbers and email addresses, even when each sequence appeared in one document of the training data [23].

Some sellers carry deletion duties. Since 1 August 2026, California data brokers "must access the “accessible deletion mechanism” at least once every 45 days and process consumer deletion requests, subject to limited exceptions" [24]. From 1 January 2028, and every 3 years after, they must undergo an audit by an independent third party [24]. California's SB 361 also requires data brokers to disclose whether they have shared data with "developers of generative AI systems", among others [24]. One cloud marketplace bars information that can identify a person unless it is publicly available, outside its extended provider program [25].

6. Quality on a sample

Audits of public corpora show the defects a sample can reveal.

Duplicates. Near-duplicates in language modeling datasets meant that "over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data" [26]. One 61-word sentence was repeated over 60,000 times in C4, and deduplicated data gave models that "emit memorized text ten times less frequently and require fewer train steps to achieve the same or better accuracy" [26].

Leakage. Train-test overlap "affects over 4% of the validation set of standard datasets" [26]. In C4, "4.6% and 5.7% of the examples" of two evaluation sets appeared verbatim, and the share of GLUE test inputs found ranged "from less than 2% to over 50%" [27].

Generated text. C4 held machine-generated text and examples from other benchmarks, and its blocklist filtering "disproportionately removes text from and about minority individuals" [27]. Indiscriminate training on model-generated content "causes irreversible defects in the resulting models, in which tails of the original content distribution disappear" [30]. The finding concerns recursive, indiscriminate use, and it is a reason to ask what share of a dataset a model produced.

Label errors. Ten benchmark test sets averaged "at least 3.3% errors", and label errors made up "at least 6% of the ImageNet validation set" [28].

Unusable text. Of "205 language-specific corpora released with five major public datasets", "At least 15 corpora have no usable text", and "these issues are easy to detect even for non-proficient speakers" [29].

What a clean sample proves. With zero defects in N records, NIST's exact binomial upper limit p solves $(1 - p)^N = \alpha/2$, so the two-sided 95% limit is $1 - 0.025^{(1/N)}$ [31].

Records checked, zero defects found	Upper limit on the true defect rate, 95% (derived) [31]
100	3.62%
200	1.83%
300	1.22%
500	0.74%
1,000	0.37%

The bound assumes the binomial model behind the method: a random sample from a large population or lot [31]. A sample the seller picked answers a different question.

7. Downstream disclosure rules

California's AB 2013 (Chapter 817, Statutes of 2024) requires developers of generative systems released on or after 1 January 2022 and made available to Californians to post documentation of the data used to train them, on or before 1 January 2026 and before each later release or substantial modification [33]. Read as enacted in 2024, its items include [33]:

- "(1) The sources or owners of the datasets."
- "(5) Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain."
- "(6) Whether the datasets were purchased or licensed by the developer."
- "(7) Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140."
- "(10) The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing."
- "(12) Whether the generative artificial intelligence system or service used or continuously uses synthetic data generation in its development."

Training "includes testing, validating, or fine tuning by the developer" [33]. A developer that buys data needs the seller's answers to items like these.

The EU AI Act requires providers of general-purpose models to keep a copyright policy that honors rights reservations under Article 4(3) of Directive (EU) 2019/790, and to publish "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" [15]. Their technical documentation covers "how the data was obtained and selected" [15]. These duties apply from 2 August 2025, and providers of models placed on the market before then must comply by 2 August 2027 [15].

The template sorts data by how it was obtained [16]. Section 2.2.1 covers datasets under "transactional commercial licensing agreements" concluded with rightsholders, and section 2.2.2 covers "other private datasets obtained through data intermediaries" without such licenses [16]. Licensed data needs "limited disclosure", while private third-party datasets are listed only if publicly known and otherwise described by modality, nature of content and language [16]. For scraped data the template asks for "the crawlers used, their purpose and behaviour, the period of collection" and the sources [16]. A provider in scope needs to know whether its seller is the rightsholder or its representative, which decides the section a purchased dataset falls in [16].

8. Checks after onboarding

SEC examination staff observed in April 2022 advisers "that had an onboarding process for alternative data service providers, but did not have a system for determining when due diligence needed to be re-performed based on passage of time or changes in data collection practices" [34]. Staff also found "ad hoc and

inconsistent diligence of alternative data service providers", and some "advisers did not apply their due diligence process to all sources of alternative data" [34]. The alert "represents the views of the staff" and "is not a rule, regulation, or statement of the Securities and Exchange Commission" [34].

FISD's identification factors note that firms may ask vendors to update their diligence questionnaires periodically, and name an "Uncooperative/unresponsive vendor" and "Unexplained inconsistencies" between diligence materials and vendor statements as red flags [35]. In 2021 the SEC charged an alternative data provider that told companies "their data would be aggregated and anonymized before being used by a statistical model", while it "used non-aggregated and non-anonymized data to alter its model-generated estimates" [32]. It was ordered to pay a $10 million penalty and settled without admitting or denying the findings [32].

Versions supply the trigger. A new MAJOR Croissant version marks changed records or an incompatible extension, and isLiveDataset marks data that keeps changing [3]. Datasheets ask whether a dataset is updated and whether older versions stay supported [1].

9. The checklist in one table

Check	Ask the seller for	What it guards against	Sources
Documentation	A datasheet covering collection, composition, distribution and maintenance, with a license URL and a version	A listing that leaves out restrictions and changes	[1, 2, 3, 4, 25]
Upstream rights	The original sources, their licenses and terms, and the upstream agreements	Platform labels and dataset licenses that understate source restrictions	[6, 7, 35]
Acquisition	How copies were acquired, crawl dates, and robots.txt and paywall handling	Pirated or restricted sources	[8, 13, 14, 17]
Model and output rights	Which of research, internal use, output sale and model release the license grants, and what termination requires	A trained model the license does not let you use	[9, 10, 11, 12]
Generated content	Which parts a model produced, and under which service terms	Terms of use and recursive-training defects	[6, 30]
Personal data	Fields with personal data, the privacy promise at collection, consent records and the de-identification method	Model deletion orders and re-identification	[18, 19, 20, 22]

Check	Ask the seller for	What it guards against	Sources
Deletions	How deletion requests and corrections reach delivered copies	Seller deletion duties that delivered copies may not reflect	[1, 24]
Sample quality	A random sample of agreed size, checked for duplicates, leakage, label errors and unusable text	Defects a listing does not state	[26, 27, 28, 29, 31]
Disclosure facts	Source owners, collection period, copyright and personal data status, and whether the seller is the rightsholder or its representative	Gaps in the buyer's own disclosures	[15, 16, 33]
Re-checks	A trigger to repeat diligence on new versions and changed collection practices	Diligence that goes stale	[3, 34, 35]

Implications

For data and machine learning leads. Write the intended use first: training from scratch, fine-tuning, evaluation, internal outputs or a shipped model, since license rights split along those lines [9, 12]. Size the random sample from the defect rate you can accept, and read part of it by hand [29, 31]. Keep the dataset version and checksum with every training run [3].

For legal and procurement. Ask for the upstream licenses and agreements, since published audits found platform labels and dataset licenses understating the restrictions on sources [6, 7], and FISD lists upstream agreements among the documents to request [35]. Read the termination clause for model artifacts [12], check whether models and outputs fall outside the license's conditions, as CDLA states for Results [10], and ask whether the license warrants title, which CDLA disclaims [10]. NIST's suggested contract terms cover ownership, usage rights, quality, security and provenance [5]. Ask for the facts your own disclosures may need, including whether the seller is the rightsholder or its representative [16, 33].

For privacy teams. Ask under which privacy promise and terms each part was collected, and when [20]. Treat de-identified data with many attributes as possibly re-identifiable [22], and remember that a model can emit personal data from its training set [23]. Ask how deletion requests reach delivered copies [24], and plan for the upstream assessment the EDPB expects of model deployers [21].

Limits of this analysis

- This note is research on what to check, never legal advice.
- It covers U.S. federal records, two California measures and the EU AI Act with its guidance, and no other jurisdiction.
- AB 2013 is read as enacted in 2024. Later amendments and its exemptions were not checked.
- It does not say when the EU AI Act's Article 10 applies, because changes to the high-risk timetable were not checked.
- The Copyright Office report is a pre-publication version [13], the court order is one district court's ruling [14], the FTC post and the SEC alert are staff views [20, 34], and the FISD papers address investment firms [12, 35].
- The license audits cover public datasets, and the 2024 audit covers fine-tuning text datasets only [6, 7]. No source here surveys warranties, indemnities, exclusivity or audit rights in paid licenses.
- The quality figures come from public corpora studied in papers from 2021 and 2022 [26, 27, 28, 29]. They show kinds of defect, never the rate in any dataset offered for sale.
- The sample bounds assume a random sample from a large population or lot [31].

Sources

1. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione and others, Communications of the ACM, Vol. 64, No. 12, December 2021, arXiv:1803.09010, version 8, 1 December 2021, viewed 7 November 2026. https://arxiv.org/abs/1803.09010

2. Croissant RAI Specification, Version 1.0. Mubashara Akhtar, Nitisha Jain, Joan Giner-Miguelez and others, MLCommons, 6 March 2024, viewed 7 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-rai-spec.html

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

4. Data Provenance Standards, v1.0.0. Data & Trust Alliance, July 2024, viewed 7 November 2026. https://dtaalliance.org/work/data-provenance-standards/

5. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. National Institute of Standards and Technology, July 2024, viewed 7 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

6. A large-scale audit of dataset licensing and attribution in AI. Shayne Longpre, Robert Mahari, Anthony Chen and others, Nature Machine Intelligence, Vol. 6, p. 975, 30 August 2024, viewed 7 November 2026. https://www.nature.com/articles/s42256-024-00878-8

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

8. 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

9. Towards Standardization of Data Licenses: The Montreal Data License. Misha Benjamin, Paul Gagnon, Negar Rostamzadeh and others, arXiv:1903.12262, 21 March 2019, viewed 7 November 2026. https://arxiv.org/abs/1903.12262

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

11. Using CC-Licensed Works for AI Training. Creative Commons, 15 May 2025, modified 28 May 2025, viewed 7 November 2026. https://creativecommons.org/using-cc-licensed-works-for-ai-training-2/

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. Copyright and Artificial Intelligence, Part 3: Generative AI Training, pre-publication version. U.S. Copyright Office, 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

14. Bartz v. Anthropic PBC, No. C 24-05417 WHA, Order on Fair Use, Document 231. U.S. District Court for the Northern District of California, 23 June 2025, viewed 7 November 2026. https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_24-cv-05417/pdf/USCOURTS-cand-3_24-cv-05417-0.pdf

15. Regulation (EU) 2024/1689 (Artificial Intelligence Act). European Parliament and Council of the European Union, 13 June 2024, Official Journal of the European Union, 12 July 2024, viewed 7 November 2026. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

16. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, AI Office, published 24 July 2025, Communication C(2025) 8311 final of 5 December 2025, viewed 7 November 2026. https://ec.europa.eu/newsroom/dae/redirection/document/118480

17. General-Purpose AI Code of Practice, Copyright chapter. European Commission, 10 July 2025, viewed 7 November 2026. https://ec.europa.eu/newsroom/dae/redirection/document/118115

18. FTC Finalizes Order with X-Mode and Successor Outlogic Prohibiting it from Sharing or Selling Sensitive Location Data. U.S. Federal Trade Commission, press release, 12 April 2024, viewed 7 November 2026. https://www.ftc.gov/news-events/news/press-releases/2024/04/ftc-finalizes-order-x-mode-successor-outlogic-prohibiting-it-sharing-or-selling-sensitive-location

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. AI (and other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive. U.S. Federal Trade Commission, Office of Technology and Division of Privacy and Identity Protection staff, 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

21. 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/system/files/documents/2024-12/edpb_opinion_202428_ai-models_en.pdf

22. Estimating the success of re-identifications in incomplete datasets using generative models. Luc Rocher, Julien M. Hendrickx and Yves-Alexandre de Montjoye, Nature Communications, Vol. 10, Article 3069, 23 July 2019, viewed 7 November 2026. https://www.nature.com/articles/s41467-019-10933-3

23. Extracting Training Data from Large Language Models. Nicholas Carlini, Florian Tramer, Eric Wallace and others, arXiv:2012.07805, version 2, 15 June 2021, viewed 7 November 2026. https://arxiv.org/abs/2012.07805

24. Information for Data Brokers. California Privacy Protection Agency, undated web page, viewed 7 November 2026. https://cppa.ca.gov/data_brokers/

25. Publishing guidelines for AWS Data Exchange. AWS Data Exchange User Guide, Amazon Web Services, undated, viewed 7 November 2026. https://docs.aws.amazon.com/data-exchange/latest/userguide/publishing-guidelines.html

26. Deduplicating Training Data Makes Language Models Better. Katherine Lee, Daphne Ippolito, Andrew Nystrom and others, ACL 2022, arXiv:2107.06499, version 2, 24 March 2022, viewed 7 November 2026. https://arxiv.org/abs/2107.06499

27. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Jesse Dodge, Maarten Sap, Ana Marasović and others, EMNLP 2021, arXiv:2104.08758, version 2, 30 September 2021, viewed 7 November 2026. https://arxiv.org/abs/2104.08758

28. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Curtis G. Northcutt, Anish Athalye and Jonas Mueller, NeurIPS 2021 Track on Datasets and Benchmarks, arXiv:2103.14749, version 4, 7 November 2021, viewed 7 November 2026. https://arxiv.org/abs/2103.14749

29. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Julia Kreutzer, Isaac Caswell, Lisa Wang and others, Transactions of the Association for Computational Linguistics, Vol. 10, 2022, arXiv:2103.12028, version 4, 21 February 2022, viewed 7 November 2026. https://arxiv.org/abs/2103.12028

30. AI models collapse when trained on recursively generated data. Ilia Shumailov and others, Nature, Vol. 631, pp. 755 to 759, 24 July 2024, viewed 7 November 2026. https://www.nature.com/articles/s41586-024-07566-y

31. NIST/SEMATECH e-Handbook of Statistical Methods, 7.2.4 Does the proportion of defectives meet requirements? and 7.2.4.1 Confidence intervals. National Institute of Standards and Technology, undated, viewed 7 November 2026. https://www.itl.nist.gov/div898/handbook/prc/section2/prc24.htm and https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

32. SEC Charges App Annie and its Founder with Securities Fraud, Press Release 2021-176. U.S. Securities and Exchange Commission, 14 September 2021, viewed 7 November 2026. https://www.sec.gov/newsroom/press-releases/2021-176

33. AB 2013, Generative artificial intelligence: training data transparency, Chapter 817, Statutes of 2024. California Legislature, chaptered 28 September 2024, text on CalMatters Digital Democracy, viewed 7 November 2026. https://calmatters.digitaldemocracy.org/bills/ca_202320240ab2013

34. Investment Adviser MNPI Compliance Issues, Risk Alert. U.S. Securities and Exchange Commission, Division of Examinations, 26 April 2022, viewed 7 November 2026. https://www.sec.gov/files/code-ethics-risk-alert.pdf

35. Alternative Data Identification Factors, Version 1.0. FISD Alternative Data Council, Software & Information Industry Association, March 2023, viewed 7 November 2026. https://fisd.net/wp-content/uploads/2023/03/FISD-Alternative-Data-Identification-Factors-11.1.22.pdf

36. The General-Purpose AI Code of Practice. European Commission, Shaping Europe's digital future, last updated 7 October 2026, viewed 7 November 2026. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

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. Buying training data: what to check before you license a dataset. 8 November 2026. https://rillor.com/insights/buying-training-data

Work with Rillor

Send the datasets you need, the coverage period, the delivery format and the intended use. Contact us at rillor.com/contact?type=market-data or contact@rillor.com. More on data marketplace: rillor.com/data/marketplace