Rillor

METHODS NOTE

Collecting training data: sources, terms and capture records

Data collection for model training as a method: name the sources, check each one's terms and opt-out signals before capture, and record where and when each item came from.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Data collection for model training holds up later when every item carries its source, its terms and its capture time, recorded at the moment of capture.

- Collection starts from a named source list. The OECD groups collection mechanisms into data collected directly from individuals and organizations and data from third-party providers [30]. The U.S. Copyright Office notes that "'Publicly available' is not synonymous with 'authorized.'" [23]
- Published license labels are often missing or wrong. An audit of more than 1,800 text datasets found license omission of 70% or more and error rates of 50% or more on widely used dataset hosting sites [3].
- RFC 9309 states that robots.txt rules "are not a form of access authorization" [1]. A 2024 audit of 14,000 web domains found robots.txt restrictions on 5 to 7% of tokens in three public web corpora in April 2024, up from about 1% in mid 2023 [2].
- Publicly accessible personal information stays subject to data protection law in most jurisdictions, according to a 2023 joint statement of data protection authorities [22]. The European Data Protection Board (EDPB) and France's data protection authority, the CNIL, list measures to take before collection, such as criteria set in advance and the exclusion of websites that object [17, 20].
- A capture record can follow the WARC web archive standard, which gives each captured response its target URL, every record a UTC capture time, and an optional content digest [5].
- Sources change after capture. Pew Research Center found that a quarter of the webpages that existed at some point between 2013 and 2023 were no longer accessible in October 2023 [25].
- Evaluation data needs its own dates. One study found train-test overlap in over 4% of the validation set of standard datasets [26], and California's training data disclosure law, AB 2013, defines training to include "testing, validating, or fine tuning" [16].
- Disclosure rules ask for what a capture log holds: AB 2013 for the sources, licensing and collection period of training datasets [16], and the EU's public summary template for crawler names, the collection period and top domain names [15].

Background

Data collection for model training covers the training data a model learns from and the evaluation data it is tested on. Its sources run from crawl archives and public bulk files to licensed collections and an organization's own systems [23, 30].

Common Crawl, a nonprofit founded in 2007, reports over 300 billion pages spanning 15 years and 3 to 5 billion new pages added each month [7]. Corpora are built from its snapshots: C4 from one [27], and FineWeb, a 15-trillion token dataset, from 96 [38]. After filtering and merging, an item's origin can be hard to rebuild from outside: the Internet Archive had indexed about 65% of the URLs in C4's English portion [27], so an archive cannot always show what a page said when it was collected.

Three developments make the record at capture matter. Robots.txt restrictions on three public web corpora rose from about 1% of tokens in mid 2023 to 5 to 7% in April 2024 [2]. Data protection authorities have published guidance on collecting personal data from the web [17, 18, 20, 22]. Disclosure rules ask for summaries of training data: California's AB 2013 from 1 January 2026 [16], and the EU's public summary of training content from 2 August 2025 [15].

This note is for teams who plan, run or review a collection. It is research, never legal advice.

Method

What was examined. Each step of a collection, from naming sources to the records a disclosure needs, read against the standards, studies, regulators' guidance and legal texts that define or measure it.

Sources. Thirty-nine sources, each opened on 7 November 2026 and read for the passages cited: 8 standards and format descriptions, 9 legal and official texts or guidance, 7 statements and guidance documents from data protection authorities and FTC staff, 8 research papers and preprints, studies from Pew Research Center and the OECD, and 5 pages of source documentation from Common Crawl and Wikimedia.

How. Each figure is taken as its source publishes it, with its date. Where a primary legal text refused automated access, the note cites an official restatement or a public copy of the text: Article 4(3) of Directive (EU) 2019/790 through the AI Act Service Desk, the Code of Practice and the CNIL [11, 13, 20], GDPR points through the EDPB and the CNIL [17, 18, 20, 21], AB 2013 through its chaptered text as published by CalMatters Digital Democracy [16], and the ISO WARC standard through the IIPC specification and the Library of Congress [5, 6]. No crawler other than Common Crawl's is named.

Where sources disagree. The CNIL lists ai.txt and TDMRep as signals beyond robots.txt and notes that "there are no standards in this area" [20], while an IETF working group is chartered to standardize such signals [36]. The 2024 opt-out audit notes that its estimates of how many head domains had no robots.txt or no terms of service page in 2016 "may be slightly high", especially for terms of service pages, because of gaps in the Wayback Machine [2].

Findings

1. A named source list, sorted by collection mechanism

Mechanisms. The OECD's October 2025 paper, Mapping relevant data collection mechanisms for AI training, divides collection into data collected directly from individuals and organizations, such as provided and observed data and voluntary data donations, and data from third-party providers, such as commercial data licensing, open data and data scraping [30]. It notes that "Data marketplaces and data brokers play a relevant role as data intermediaries" [30].

Routes in practice. The U.S. Copyright Office lists "publicly available" data downloaded from the internet, pre-existing databases such as "an entire copy of Wikipedia using one of the regularly provided backups", text gathered by web crawlers and "often obtained from Common Crawl", and data "from licensed or non-public sources" [23]. Its footnote adds that "'Publicly available' is not synonymous with 'authorized.' It may simply be used to mean 'available on the internet.'" [23]

Crawl archives. Common Crawl publishes WARC files "which store the raw crawl data", WAT files with computed metadata and WET files with extracted plain text [9]. A corpus drawn from it inherits its snapshot dates. C4 applies a set of filters to a single snapshot [27], and FineWeb's authors document their deduplication and filtering choices across 96 snapshots [38].

Naming the mechanism. Croissant RAI 1.0, published by MLCommons on 6 March 2024, gives a dataset a `rai:dataCollectionType` whose recommended values include Document analysis, Software Collection, Web Scraping and Web API, with `rai:dataCollectionRawData` for the source of the data and `rai:dataCollectionTimeframe` for the start and end dates [31]. Datasheets for Datasets asks creators to read its collection questions "prior to any data collection to flag potential issues" [4]. They cover the mechanisms used, the sampling strategy, the timeframe, and whether the data came from individuals directly or from third parties such as websites [4].

Creation time and capture time. Datasheets asks whether the collection timeframe matches the creation timeframe of the data, as in a "recent crawl of old news articles" [4]. In C4, of 1,000,000 randomly sampled URLs, 92% are estimated to have been written between 2011 and 2019, with "a non-trivial amount of data that was written between 10-20 years before data collection" [27]. A source list that records both dates shows the gap.

2. Terms recorded per item, in layers

Within one source. Wikimedia licenses all original dump text under the GFDL and the Creative Commons Attribution-Share-Alike 4.0 License, but "many images are NOT released under this license", and each image's copyright information sits on its description page inside the dumps [35]. Wikinews text from 25 September 2005 is under Creative Commons Attribution 2.5, and Wikidata's structured data is waived under CC0 [35]. The page calls itself "a high-level guide only" and defers to the terms of use and the information inside the dumps [35]. The license is a field of the item.

Over a crawl archive. Common Crawl's terms of use, last updated 7 March 2024, say crawled content "may be subject to separate terms of use or terms of service from the owners of such Crawled Content", and users agree to respect the copyrights and other rights of third parties in it [10]. Common Crawl may amend the terms at any time [10], so a capture record keeps the date of the version in force.

Government works. Under 17 U.S.C. 105, copyright protection "is not available for any work of the United States Government", but the government may hold copyrights transferred to it "by assignment, bequest, or otherwise" [24]. Items on one agency site can carry different terms.

Labels on published datasets. The Data Provenance Initiative audited more than 1,800 text datasets and found "license omission of 70%+ and error rates of 50%+" on widely used dataset hosting sites [3]. Its re-annotation found 66% of the analyzed licenses on one hosting site "in a different use category, often labeled as more permissive than the author's intended license" [3].

License detection in code. The Stack's builders started from 220.92 million repository names in GHArchive events from 1 January 2015 to 31 March 2022 [28]. GHArchive supplied licenses for 26.4 million repositories, and for the rest a license detector "did not detect a license for more than 80% of the repositories" [28]. The dataset offers a tool, "Am I in The Stack", and a process for code to be removed [28].

Lawful access. In the Copyright Office's view, "the knowing use of a dataset that consists of pirated or illegally accessed works should weigh against fair use without being determinative" [23]. It lists the copies a collection makes: downloads, transfers across storage media, format conversions, and modified versions or filtered subsets [23].

First-party data. In a staff blog post of 13 February 2024, FTC staff wrote that a business "cannot then unilaterally renege on those commitments after collecting users' data", and that adopting more permissive data practices through a "surreptitious, retroactive amendment" to its terms or privacy policy may be unfair or deceptive [37]. Data from an organization's own systems carries the notices it was collected under.

3. Opt-out signals: robots.txt, the EU reservation and a preference vocabulary

What robots.txt is. RFC 9309, on the Standards Track since September 2022, defines rules "that crawlers are requested to honor when accessing URIs" and states that "These rules are not a form of access authorization" [1]. It calls the protocol "not a substitute for valid content security measures" [1].

Status codes decide. After a 4xx answer for robots.txt, "the crawler MAY access any resources on the server" [1]. After a 5xx answer, "the crawler MUST assume complete disallow", and after "a reasonably long period of time (for example, 30 days)" it may treat the file as unavailable or keep using a cached copy [1]. Crawlers "SHOULD NOT use the cached version for more than 24 hours, unless the robots.txt file is unreachable" [1]. The status of each robots.txt fetch is part of what a capture saw.

Who fetched. Common Crawl's crawler identifies itself as CCBot/2.0 [8]. Common Crawl says it is "aware of crawlers falsely identifying themselves as CCBot" and runs CCBot on dedicated IP address ranges with reverse DNS [8]. A user agent string alone does not prove who fetched a page.

The EU reservation. Article 53(1)(c) of the EU AI Act requires providers of "general-purpose AI models" to "put in place a policy to comply with Union law on copyright and related rights", including to identify and comply with a reservation of rights under Article 4(3) of Directive (EU) 2019/790 [11]. The Code of Practice's Copyright Chapter restates that rightsholders may reserve text and data mining "in any appropriate manner, such as machine-readable means in the case of content made publicly available online" [13]. The CNIL adds that this "includes metadata and terms and conditions of a website or a service" [20].

The Code's crawling commitments. The Code, published on 10 July 2025, is "a voluntary tool" [12]. Its signatories commit to crawlers "that read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt)" as RFC 9309 specifies, to other appropriate machine-readable protocols, to respect for paywalls and subscription restrictions, and to publishing information about their crawlers with a way to be notified of updates [13].

How often signals change and disagree. The Consent in Crisis audit covered 14,000 web domains: a head sample of the top 2,000 domains by token count in each corpus and a random sample of 10,000 [2]. Robots.txt restrictions covered about 1% of C4, RefinedWeb and Dolma tokens in mid 2023 and 5 to 7% in April 2024, and 20 to 33% of tokens in the head sample, up from under 3% [2]. Terms of service restrictions covered 45% of C4 [2]. Many sites "write anti-crawling terms but have no robots.txt file (35.1%), or have no ToS but a restrictive robots.txt (20.3%)" [2].

A vocabulary for preferences. The IETF's AI Preferences working group is chartered to standardize ways to express "preferences about how content is collected and processed", through content metadata or the delivery protocol, robots.txt and HTTP header fields included, with technical enforcement out of scope [36]. Its vocabulary, draft-ietf-aipref-vocab-08, last updated 13 September 2026, defines a training category labeled `train-ai`, and each category resolves to "allowed", "disallowed" or "unknown" [36]. The charter set August 2026 as the milestone for sending the vocabulary to the Internet Engineering Steering Group (IESG) [36]. On 7 November 2026 the datatracker shows the document as an Internet-Draft in IESG state "I-D Exists" [36].

4. Personal data in collected sources

Public pages sit inside privacy law. The 2023 joint statement on data scraping says that "Personal information that is publicly accessible is still subject to data protection and privacy laws in most jurisdictions" [22]. The October 2024 concluding statement, with 16 co-signatories in all, says organizations that train models on scraped data sets or on data from their own platforms "must comply with data protection and privacy laws as well as any AI-specific laws where those exist" [22].

Measures before collection (EDPB). The EDPB's Opinion 28/2024, adopted on 17 December 2024, notes that scraped content from news outlets, social media, forums and personal websites "may contain personal data" [17]. Its measures "may include" leaving out certain data categories or sources, excluding websites that clearly object to web scraping, "for example, by respecting robots.txt or ai.txt files", and limits such as time periods [17]. It also describes an opt-out list that lets people object to collection from certain websites "including before the data collection occurs" [17].

Recording and validating (EDPB guidelines). Guidelines 03/2026 on web scraping, version 1.0, were adopted for public consultation on 7 July 2026, and the consultation ran until 30 October 2026 [18, 19]. They recommend official, maintained sources over "secondary or outdated aggregators", a record of "when the data was scraped", and validation of the data before use [18]. They call it good practice to include information on "the domain names and URLs of the scraped web pages" with "the date or period of collection" [18]. They state that the absence of a robots.txt file "does not amount to consent within the meaning of the GDPR" [18].

Mandatory measures (CNIL). The CNIL's focus sheet of 5 January 2026 says collection by web scraping "is generally based on legitimate interest" [20]. Its mandatory measures include defining "in advance, specific collection criteria", filtering out unneeded categories, deleting irrelevant data "immediately after collection or as soon as it is identified as such", and excluding websites that oppose scraping through robots.txt or CAPTCHA [20].

Keeping the source. The CNIL also advises retaining metadata about the source of collection "to facilitate the retrieval of a person or a specific piece of data within the database", such as "the display name and URL source of each collected image" [21]. The source URL in a capture record serves that retrieval.

5. The capture record, field by field

WARC. The IIPC's WARC 1.1 specification makes WARC-Record-ID, WARC-Date, WARC-Type and Content-Length mandatory [5]. WARC-Date "shall represent the instant that data capture for record creation began", as a UTC timestamp, and WARC-Target-URI is "the original URI whose capture gave rise to the information content in this record" [5]. WARC-Payload-Digest gives "the algorithm name and calculated value of a digest applied to the payload", and "No particular algorithm is recommended" [5]. The format defines eight record types, among them request, response, metadata and revisit, and keeps the harvesting protocol's request headers as well as its responses [5]. WARC was approved as an ISO International Standard in May 2009 [6].

Provenance links. W3C's PROV-O, a Recommendation of 30 April 2013, defines prov:generatedAtTime, prov:wasDerivedFrom, prov:hadPrimarySource, prov:wasRevisionOf and prov:invalidatedAtTime [32]. A capture, a revision and a withdrawal can each be stated with them.

Dataset fields. The Data & Trust Alliance's Data Provenance Standards "consist of 22 metadata fields, grouped into 3 standards: Source, Provenance, and Use", among them the dataset issue date, the date of the previously issued version, the method and the license to use [33]. NIST AI 600-1 suggests that system inventories include "Data provenance information (e.g., source, signatures, versioning, watermarks)" [34].

Raw data and maintenance. Datasheets asks whether the "raw" data was saved in addition to the cleaned data, how often the dataset is updated and whether older versions stay available [4].

The fields together. Each row names the source behind it.

Field	What it records	Basis
Record ID and type	A unique ID and the record type, such as response, revisit or metadata	[5]
Source URL	The URI the crawler requested, kept searchable	[5, 18, 21]
Capture time	The UTC instant capture began	[5, 18]
Content digest	The digest algorithm and value for the payload	[5]
Fetch status	The HTTP status of the page and of robots.txt	[1, 5]
Opt-out signals	The robots.txt rules and other signals read at capture	[1, 13, 17, 20]
Terms	The source's terms and the item's license, with their dates	[10, 35]
Personal data	Whether the item holds personal information	[16, 17]
Collection period	Start and end dates of the collection	[15, 16, 31]
Version	The earlier capture this one revises	[32, 33]

6. Sources change after capture

Link rot. Pew Research Center sampled about 90,000 pages a year collected by Common Crawl from 2013 to 2023 [25]. A quarter of the pages that existed at some point in that span were no longer accessible in October 2023, 16% from root domains that still worked and 9% because the whole root domain had gone [25]. Some 38% of pages from 2013 were unavailable, against 8% of pages from 2023, and 11% of all references linked on Wikipedia were no longer accessible [25].

Lists go stale. Of the 220.92 million repository names The Stack started from, 137.36 million were downloaded, "a clone success rate of over 62%", because some had been deleted or made private [28].

Signals change. Robots.txt restrictions moved from about 1% to 5 to 7% of tokens in three public corpora between mid 2023 and April 2024 [2], so a re-capture can meet a different answer from the first. Common Crawl's own terms carry their revision date [10].

Archives fill some gaps. The Memento framework, RFC 7089, "bridges the present and past Web" with Accept-Datetime and Memento-Datetime headers, so a client can ask an archive for a page as it stood at a given time [39]. Archives are partial: the Internet Archive had indexed about 65% of the URLs in C4's English portion [27].

Versions. WARC's revisit record, with its "Identical Payload Digest" and "Server Not Modified" profiles, logs a re-capture whose content did not change [5]. When content changes, prov:wasRevisionOf links the new version to the old one [32], and the Data & Trust Alliance's field for the date of the previous version does the same at dataset level [33].

7. Evaluation data kept apart

Overlap. Lee and colleagues found that "over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data", removed from C4 "a single 61 word English sentence that is repeated over 60,000 times", and measured "train-test overlap, which affects over 4% of the validation set of standard datasets" [26]. The Stack's authors report that "near-deduplicating the data significantly boosts performance across all experiments" [28].

Benchmarks inside a corpus. The C4 documentation study found that "the percentage of inputs found in C4.EN varies widely, from less than 2% to over 50%" across the benchmark datasets it examined [27].

Collecting after a cutoff. LiveBench's authors call test set contamination "a well-documented obstacle for fair LLM evaluation" [29]. Their benchmark "contains frequently-updated questions from recent information sources", such as recent math competitions, arXiv papers and news articles, and questions "are added and updated on a monthly basis" [29]. A collection with creation and capture dates on every item can apply the same rule, with evaluation items dated after the training data's collection period.

Same records for both. AB 2013 says training "includes testing, validating, or fine tuning" [16], so evaluation datasets fall inside the documentation it requires.

8. What disclosure rules ask a capture log to answer

California. AB 2013, chaptered on 28 September 2024 as Chapter 817 of the Statutes of 2024, adds Title 15.2, Sections 3110 and 3111, to the Civil Code [16]. On or before 1 January 2026, and before each later release or substantial modification, the developer of "a generative artificial intelligence system or service" released on or after 1 January 2022 shall post documentation on the data used to train it, including "A high-level summary of the datasets" [16]. Section 3111(a) lists twelve items [16]. Among them are "The sources or owners of the

datasets", the number of data points, whether the datasets "were purchased or licensed", whether they include personal information, any cleaning or processing, and "The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing" [16].

European Union. Article 53(1)(d) of the EU AI Act requires providers to publish "a sufficiently detailed summary about the content used for training", on a template provided by the AI Office [11]. The Commission published the template on 24 July 2025 [14]. Its section 2.3, on data crawled and scraped from online sources, asks for crawler names, their purposes and behavior, including "respect of robot.txt and other protocols", the "Period of data collection" and the "top 10 % of all domain names determined by the size of the content scraped" [15]. Small and medium-sized enterprises should give the top 5% of domain names or 1,000 domain names, whichever is lower [15].

Dates. The notice makes the obligation to publish summaries applicable from 2 August 2025 [15]. Providers of models placed on the market before that date should publish theirs no later than 2 August 2027, and the notice dates the start of the AI Office's supervision and enforcement to 2 August 2026 [15].

Mapping. A capture log with the fields in finding 5 answers the sources, terms, personal data and collection period from its own records, and its robots.txt answers show what each crawl met.

Implications

For teams starting a collection. A collection that can answer these questions later:

1. Names every source and its collection mechanism before the first capture [4, 30, 31].
2. Records the terms in force for the source and for each item, with their dates [10, 24, 35].
3. Treats a missing license label as a gap to resolve, since published labels are often missing or wrong [3, 28].
4. Records the robots.txt answer and its HTTP status at each capture, with any other machine-readable reservation found [1, 13, 20].
5. Sets personal data criteria in advance and keeps the source URL of each item [17, 20, 21].
6. Stores each capture with its target URL, UTC capture time and content digest [5].
7. Keeps re-captures as revisits or as new versions linked to the old [5, 32].
8. Dates evaluation items and keeps them apart from training data [26, 27, 29].
9. Keeps the collection period and licensing status ready for a disclosure summary [15, 16].

For engineers. A WARC file holds request, response and metadata records in one container [5], so the capture log and the raw copy can share one format. A matching payload digest is the basis of WARC's "Identical Payload Digest" revisit profile, which records a re-capture whose content did not change [5]. Times are kept in UTC, as WARC-Date requires [5].

For privacy and rights reviewers. The documents cited here are research inputs, and this note gives no legal advice. Guidance differs by jurisdiction, and two of the texts cited are a consultation version and a pre-publication version [18, 23].

At Rillor. Rillor's data collection for model training names the sources before collection starts and checks each source's terms before anything is captured. Every record keeps its capture time, the raw copy is kept unchanged, and when a source changes, the change is kept as a new version. The client receives the dataset, its capture log, the source list with terms and its version history.

Limits of this analysis

- This note describes a collection method from published sources. It is research, never legal advice, and it ranks no tool, vendor or archive.
- The legal texts covered are Article 53 of the EU AI Act with its template and Code of Practice, California's AB 2013 and 17 U.S.C. 105, with privacy guidance from the EDPB, the CNIL, a joint statement of authorities and FTC staff. Other jurisdictions and U.S. court decisions are outside it.
- The primary texts of Directive (EU) 2019/790, the GDPR and ISO 28500 refused automated access, and are cited only through official restatements and the IIPC specification [5, 11, 13, 18, 20].
- The opt-out figures cited here are from 2023 and 2024 and cover three corpora, and the audit notes that its early estimates of domains with no robots.txt or terms of service page may be slightly high [2]. No source opened measures how often a source's terms change after capture.
- Pew measured whether pages were accessible, and its figures say nothing about pages whose content changed [25].
- The EDPB guidelines are cited in version 1.0, adopted for public consultation [18], the Copyright Office report in its pre-publication version [23], and the IETF vocabulary as an Internet-Draft [36].
- Figures from studies published between 2021 and 2024 describe the corpora they measured, and are no guide to current rates.

Sources

1. RFC 9309: Robots Exclusion Protocol. M. Koster, G. Illyes, H. Zeller and L. Sassman, IETF, Standards Track, September 2022, viewed 7 November 2026. https://www.rfc-editor.org/rfc/rfc9309.html

2. 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

3. 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

4. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione and others, Communications of the ACM, arXiv 1803.09010, version 8, 1 December 2021, viewed 7 November 2026. https://arxiv.org/abs/1803.09010

5. The WARC Format 1.1. International Internet Preservation Consortium, WARC specifications, viewed 7 November 2026. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/

6. WARC, Web ARChive file format (fdd000236). Library of Congress, Sustainability of Digital Formats, viewed 7 November 2026. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

7. Common Crawl, home page. Common Crawl Foundation, viewed 7 November 2026. https://commoncrawl.org/

8. CCBot. Common Crawl Foundation, viewed 7 November 2026. https://commoncrawl.org/ccbot

9. Get Started. Common Crawl Foundation, viewed 7 November 2026. https://commoncrawl.org/get-started

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

11. Article 53: Obligations for providers of general-purpose AI models, Regulation (EU) 2024/1689. European Commission, AI Act Service Desk, viewed 7 November 2026. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-53

12. The General-Purpose AI Code of Practice. European Commission, Shaping Europe's digital future, code published 10 July 2025, viewed 7 November 2026. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

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

14. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, 24 July 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

15. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models, Communication from the Commission C(2025) 8311 final. European Commission, 5 December 2025 (first published 24 July 2025 as the annex to C(2025) 5235 final), viewed 7 November 2026.

https://ec.europa.eu/newsroom/dae/redirection/document/118480

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

17. 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

18. Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0. European Data Protection Board, adopted for public consultation 7 July 2026, viewed 7 November 2026. https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf

19. EDPB sheds light on anonymisation and web scraping for generative AI and adopts final version of guidelines on blockchain. European Data Protection Board, 8 July 2026, viewed 7 November 2026. https://www.edpb.europa.eu/news/edpb-sheds-light-on-anonymisation-and-web-scraping-for-generative-ai-and-adopts-final-version_en

20. The legal basis of legitimate interest: focus sheet on the measures to implement in the case of data collection by web scraping. CNIL, 5 January 2026, viewed 7 November 2026. https://www.cnil.fr/en/legal-basis-legitimate-interest-focus-sheet-measures-implement-case-data-collection-web-scraping

21. Relying on the legal basis of legitimate interests to develop an AI system. CNIL, viewed 7 November 2026. https://www.cnil.fr/en/relying-legal-basis-legitimate-interests-develop-ai-system

22. Joint statement on data scraping and the protection of privacy, 24 August 2023, and Concluding joint statement on data scraping and the protection of privacy, October 2024. Office of the Privacy Commissioner of Canada and co-signing data protection authorities, viewed 7 November 2026. https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2023/js-dc_20230824/ and https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2024/js-dc_20241028/

23. 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

24. 17 U.S.C. 105, Subject matter of copyright: United States Government works. U.S. Copyright Office, Title 17, Chapter 1, viewed 7 November 2026. https://www.copyright.gov/title17/92chap1.html

25. When Online Content Disappears. Athena Chapekis, Samuel Bestvater, Emma Remy and Gonzalo Rivero, Pew Research Center, 17 May 2024, viewed 7 November 2026. https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/

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. The Stack: 3 TB of permissively licensed source code. Denis Kocetkov, Raymond Li, Loubna Ben Allal and others, arXiv 2211.15533, 20 November 2022, viewed 7 November 2026. https://arxiv.org/abs/2211.15533

29. LiveBench: A Challenging, Contamination-Limited LLM Benchmark. Colin White, Samuel Dooley, Manley Roberts and others, ICLR 2025, arXiv 2406.19314, version 2, 18 April 2025, viewed 7 November 2026. https://arxiv.org/abs/2406.19314

30. Mapping relevant data collection mechanisms for AI training. OECD, OECD Artificial Intelligence Papers No. 48, October 2025, viewed 7 November 2026. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/10/mapping-relevant-data-collection-mechanisms-for-ai-training_62921889/3264cd4c-en.pdf

31. Croissant RAI Specification, version 1.0. MLCommons, 6 March 2024, viewed 7 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-rai-spec.html

32. PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013, viewed 7 November 2026. https://www.w3.org/TR/prov-o/

33. Data Provenance Standards, v1.0.0. Data & Trust Alliance, launched 30 November 2023, page updated 9 July 2024, viewed 7 November 2026. https://dataandtrustalliance.org/work/data-provenance-standards

34. 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

35. License information about Wikimedia dump downloads. Wikimedia Foundation, viewed 7 November 2026. https://dumps.wikimedia.org/legal.html

36. AI Preferences (aipref) working group charter, and draft-ietf-aipref-vocab-08, A Vocabulary For Expressing AI Usage Preferences, by Paul Keller and Martin Thomson. IETF, last updated 13 September 2026, viewed 7 November 2026. https://datatracker.ietf.org/wg/aipref/about/ and https://datatracker.ietf.org/doc/draft-ietf-aipref-vocab/

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

38. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal and others, arXiv 2406.17557, version 2, 31 October 2024, viewed 7 November 2026. https://arxiv.org/abs/2406.17557

39. RFC 7089: HTTP Framework for Time-Based Access to Resource States (Memento). H. Van de Sompel, M. Nelson and R. Sanderson, IETF, Informational, December 2013, viewed 7 November 2026. https://www.rfc-editor.org/rfc/rfc7089.html

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. Collecting training data: sources, terms and capture records. 8 November 2026. https://rillor.com/insights/collecting-training-data

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data