Rillor

RESEARCH REPORT

What training data costs, from filings and studies

How much training data costs, from filings and published studies: dated figures for licensed content, human labels and model-made labels, what each covers and what it leaves out.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Public filings, court records and published studies give dated figures for what training data costs, and each figure covers a different part of the bill.

- No filing reviewed here discloses a price per record, token, image or hour of licensed content. Reddit's Form S-1 reported data licensing arrangements from January 2024 with an aggregate contract value of $203.0 million over two to three years, and named no buyer [1].
- Studies from 2021 to 2023 put human labels for text tasks at about $0.07 to $0.30 each (derived from the costs they paid or estimated), and a 2023 benchmark estimated an average payment of about $95 an hour for the experts who wrote and validated its questions [18, 19, 24, 25].
- Model-made labels cost a fraction of human labels at the model prices of 2021 to 2023: less than $0.003 per annotation in a 2023 PNAS study, and 50% to 96% less for the same downstream performance in a 2021 study [23, 24].
- Free public data still costs compute to prepare. Building FineWeb, a 15-trillion-token text dataset drawn from 96 snapshots of the Common Crawl web archive, took an estimated 80,000 H100 GPU hours of training runs to choose its filters, and scoring it for educational quality took 6,000 more [12].
- A court-approved class settlement between a model developer and authors and publishers, given final approval on 20 July 2026, pays approximately $3,000 per work before costs and fees, and its release grants no license to train on the works [8, 9].
- Amazon Mechanical Turk, the crowd work marketplace that two of the label cost studies reviewed here used for their human baseline, closed permanently on 30 September 2026 [14, 18, 23].
- Licenses for content are term contracts. Reddit's remaining performance obligations in contracts longer than a year, mostly long-term content licensing contracts, fell from $143.7 million at the end of 2025 to $62.4 million at 30 September 2026, and its filing says the agreements behind substantially all of its licensing contract value are approaching the end of their current terms [2, 3].
- A projection published in June 2024 dates the point where training sets reach the stock of public human text to between 2026 and 2032, with a median year of 2028 [33].

Background

Buyers of training data have three routes. They can license content from the people who hold it, pay for labels on data they already have, or commission a build that combines sources, labels and cleaning into one dataset.

The Copyright Office's report on generative training describes the licensing route. Developers "may also license data from third parties, such as traditional publishers, intermediaries, and specialized data providers", and may find such material desirable "because it may not be available to competitors, is reliably high-quality, or promotes particular characteristics during training" [7]. It also notes: "Because processing data on a massive scale is resource intensive, some developers rely, in whole or in part, on datasets that were initially collected and curated by third parties" [7].

Free public data sits underneath much of this. Common Crawl's site, viewed on 8 November 2026, describes Common Crawl as "a free, open repository of web crawl data that can be used by anyone", with over 300 billion pages spanning 15 years and 3 to 5 billion new pages added each month [10]. A Mozilla Foundation study from February 2024 found that at least 64% of 47 LLMs for text generation published between 2019 and October 2023, 30 models, used at least one filtered version of Common Crawl for pre-training [11].

This report covers the data. Published estimates of what it costs to train a model answer a different question. Cottier et al., in a paper revised in February 2025, count "hardware, energy, cloud rental, and staff expenses" and state that their methods "neglect several costs that are potentially significant, including the data center infrastructure apart from the training cluster, and the acquisition of data for model training" [13].

Licensing figures come from sellers' filings, which report what sellers received. Label costs come from research papers, which report what one team paid, or estimated it would pay, for one task. The Method sets out what each kind of figure shows.

Method

What was examined. Dated public figures for what training data costs a buyer, or what a seller received for it: licensed content, human labels, model-made labels, the preparation of free data and the upkeep that continues after delivery.

Sources. Thirty-six sources, each opened on 8 November 2026: SEC filings by Reddit [1, 2, 3], Shutterstock [4], John Wiley & Sons [5], CuriosityStream [6] and Innodata [29], the Copyright Office's report and index page [7, 36], one settlement record [8, 9], three platform pages [10, 14, 17], one company announcement [28], 18 research papers [12, 13, 15, 18, 19, 20, 21, 22, 23, 24, 25, 27, 30, 31, 32, 33, 34, 35] and three reports and posts, from the ILO, Stanford's Alpaca team and the Mozilla Foundation [11, 16, 26].

Evidence classes. Each figure is classed by the evidence behind it.

Evidence class	What it shows	What it leaves out	Sources
Contract value in a filing	What buyers committed to pay one seller over a term	Units, buyers and how each buyer uses the data	[1]
Revenue line or backlog in a filing	What one seller recognized, or has left to recognize, in a period	Units, buyers and other products sold in the same line	[2, 3, 4, 5, 6, 29]
Settlement payment	What a court approved per work for past conduct	Any license or right to train	[8, 9]
Posted platform rate	The pay floor and fee a platform sets for human time	Time per label, review and rework	[17]
Reported or estimated cost per label	What one team paid, or estimated it would pay, for one task at one date	Other tasks, later prices, setup and review	[18, 23, 24, 25]
Expert payment schedule	What one benchmark paid, or estimated it paid, experts per hour, per question or by prize	Labels outside evaluation sets	[19, 20]
Compute hours	Hardware time spent filtering or scoring free data	A price for those hours	[12]
Valuation	What one investment implied a whole company is worth	Any price of data	[28]

Evidence class	What it shows	What it leaves out	Sources
Projection	When a modeled trend reaches a modeled limit	Changes to the trend after publication	[33]

How. Figures are taken from the passage or table cited, in the currency and units the source uses. Figures marked "derived" are this report's arithmetic on a source's numbers. No figure is adjusted for inflation or converted between currencies. Model-made label costs are tied to the model prices of each study's year, 2021 to 2023 [23, 24, 25]. The filings name no buyers, and none are added from press reports. Press figures relayed in the Copyright Office report's footnotes are left out, because no filing reviewed states them. This report is research and gives no legal advice.

Findings

1. Licensed content in filings

The filings give contract values and revenue lines. None gives a price per unit of content.

Reddit. The Form S-1 of 22 February 2024 states: "In January 2024, we entered into certain data licensing arrangements with an aggregate contract value of $203.0 million and terms ranging from two to three years." Reddit expected "a minimum of $66.4 million of revenue to be recognized during the year ending December 31, 2024" [1]. The filing describes the business as "allowing third parties to access, search, and analyze data on our platform", delivered "through continuous access to our data API as well as quarterly transfers of Reddit data over the term of the arrangement" [1]. Spread evenly, $203.0 million over two to three years is about $68 million to $102 million a year (derived) [1]. The filing names no counterparty and does not split the value by use.

The 10-K for 2025, filed 6 February 2026, reports "other revenue" of $140.0 million in 2025, $114.7 million in 2024 and $15.2 million in 2023 [2]. Other revenue "consists of revenue from content licensing and products sold directly to users", so it is wider than licensing [2]. The same filing states that "substantially all of the contract value associated with our licensing revenue is derived from two of our partners" [2].

Shutterstock. The 10-K for 2025, filed 17 February 2026, reports Data, Distribution, and Services revenue of $203.3 million in 2025, $175.3 million in 2024 and $137.3 million in 2023, against total revenue of $989.9 million in 2025 [4]. That is 20.5% of total revenue in 2025 (derived) [4]. The line covers licenses to metadata through the company's data offering, distribution and advertising services from Giphy, Shutterstock Studios and other tailored services [4]. The filing reports "increased demand for access to our metadata for machine learning and generative artificial intelligence model training" [4], and reports no training revenue on its own. The line's results "fluctuate from quarter to quarter based on the timing of delivery of large metadata sets" [4]. Contributors "typically earn a royalty each time their work is licensed", so the seller pays for content as it licenses it [4].

John Wiley & Sons. The 10-K for the year to 30 April 2026, filed 24 June 2026, reports "Research AI licensing revenue" of $33.1 million, against approximately $11 million the year before, and "Learning AI licensing revenue" of $16.0 million, against approximately $29 million [5]. Together that is about $49.1 million in fiscal 2026 against about $40 million in fiscal 2025 (derived), with the two segments moving in opposite directions [5]. Wiley licenses this content "generally at a fixed transaction price, with revenue recognized at a point in time or a combination of point in time and over time" [5]. In fiscal 2026 it began offering licensing as a service to other publishers, combining their content with its own for LLM or corporate use in return for a royalty [5].

CuriosityStream. The 10-K for 2025, filed 12 March 2026, reports revenue of $71.7 million in 2025 against $51.1 million in 2024, growth "primarily driven by an increase of $25.4 million in Content Licensing revenue" [6]. The company describes delivering "hundreds of thousands of hours of video and audio" to technology companies for model training [6]. The filing gives volume in hours and no price per hour.

Seller	Filing and date	Figure	Evidence class	What the figure includes
Reddit	S-1, 22 February 2024	$203.0 million over two to three years [1]	Contract value in a filing	Access, search and analysis, with training among the listed purposes [1]
Reddit	10-K, 6 February 2026	$140.0 million other revenue in 2025 [2]	Revenue line or backlog in a filing	Content licensing and products sold directly to users [2]
Shutterstock	10-K, 17 February 2026	$203.3 million in 2025 [4]	Revenue line or backlog in a filing	Metadata licensing, Giphy, Studios and other services [4]
Wiley	10-K, 24 June 2026	$33.1 million and $16.0 million in two segments, fiscal 2026 [5]	Revenue line or backlog in a filing	Content licensing the filing labels "AI licensing" [5]
CuriosityStream	10-K, 12 March 2026	$25.4 million increase in Content Licensing revenue in 2025 [6]	Revenue line or backlog in a filing	Content licensing, which also includes traditional media licensing [6]

None of the five filings names a buyer or a price per record, token, image, hour or title [1, 2, 4, 5, 6].

2. Licensing in the Copyright Office and court record

As of 8 November 2026 the Copyright Office lists Part 3 of its report, on generative training, only in its pre-publication version of 9 May 2025 [7, 36].

The report records licensing activity in images, audio-visual works, academic and nonfiction publishing and news [7]. Some commenters argued "that licensing such data would be prohibitively expensive and available only to certain developers and for certain copyrighted works" [7]. Licensors "touted their products as attractive to AI companies because they can provide data that is newly released, high-quality, curated, and clean" [7].

On price structure, the report records a commenter's suggestion that "if AI companies struggle to compensate rightsholders in the near-term, rightsholders can negotiate licenses that forgo up-front payments or traditional royalties in exchange for later shares in revenues" [7]. In its analysis and recommendations on licensing, the Office finds that "voluntary licensing may be workable, at least in certain contexts", particularly "where training is focused on valuable content that can be licensed in relatively high volumes (e.g., popular music and stock photography), or in fields where the number of copyright owners is

limited" [7]. It adds that "compensation structures based on a percentage of revenue or profits, without large up-front cash outlays, may be an attractive alternative for smaller developers" [7]. A license can therefore be priced as a fixed sum, as Wiley reports [5], or as a share of later revenue [7].

A settlement for unlicensed acquisition. A class action claimed that a model developer infringed copyrights by downloading works from two allegedly pirated online datasets [9]. The settlement sets up a fund of $1.5 billion and provides "approximately $3,000 per work", before costs, fees and expenses [9]. The court granted final approval on 20 July 2026 [8]. The release "does not constitute, in any respect, a license to torrent, scan, or train AI models on any copyrighted works" [9]. The $3,000 is a payment for past conduct and prices no license. This report draws no legal conclusion from it.

3. Free public data and what preparing it costs

A free source costs nothing to download and a good deal to prepare.

The FineWeb paper (NeurIPS 2024) puts numbers on one preparation. FineWeb is "a 15-trillion token dataset derived from 96 Common Crawl snapshots" [12]. To choose its filters, the authors "trained over 70 models on our internal cluster, for an estimated total of 80,000 H100 GPU hours" [12]. For FineWeb-Edu, a model scored "460,000 randomly sampled webpages ... for their educational quality on a scale from 0 to 5", and "applying the classifier to the 15 trillion tokens of FineWeb required 6,000 H100 GPU hours" [12]. That second step is a model-made label applied at web scale.

Rights checking is a cost of free data too. The Data Provenance Initiative, in a 2023 audit, examined "1800+ text datasets" and observed "frequent miscategorization of licenses on widely used dataset hosting sites, with license omission of 70%+ and error rates of 50%+" [35].

4. Human labels and the closure of Mechanical Turk

The closure. "Amazon Mechanical Turk permanently closed on September 30, 2026" [14]. The closure "also applies to SageMaker Ground Truth and Amazon Augmented AI", and the Mechanical Turk worker type is no longer available for labeling jobs or human review workflows from that date [14].

What workers earned. Hara et al., at CHI 2018, recorded 2,676 workers performing 3.8 million tasks on Mechanical Turk and found "a median hourly wage of only ~$2/h, and only 4% earned more than $7.25/h", while "the average requester pays more than $11/h" [15]. Their figures depend on "how unpaid work is included" [15]. The International Labour Office found that across five platforms in 2017 a worker earned US$4.43 an hour of paid work on average, and US$3.31 once unpaid hours were counted [16].

A posted rate. Prolific's pricing page, viewed 8 November 2026, sets a recommended minimum reward of £9.00 or $12.00 an hour and an absolute minimum of £6.00 or $8.00, and charges a platform fee of 42.857% of rewards to corporate customers and 33.3% to academia and non-profits, with VAT on the fee only [17]. At the corporate fee, an hour at the recommended minimum costs about $17.14 (derived) [17].

Costs per label. Three studies report human label costs: AlpacaFarm paid its annotators, Wang et al. estimated from Google Cloud Platform's posted labeling price and Ding et al. estimated from interviews with linguists and professional annotators [18, 24, 25].

Study	Published	Task	Human cost as reported	Per label
Wang et al. [24]	August 2021	Language tasks including SST-2 sentiment	"$0.11 per 50 input tokens with a minimum of $0.11", estimated from Google Cloud Platform's labeling price	$0.11 as estimated
Ding et al. [25]	June 2023	SST-2 sentiment	$221 to $300 for 3,000 samples, $4,800 to $6,700 for 67,349, estimated from interviews	About $0.07 to $0.10 (derived)
Dubois et al., AlpacaFarm [18]	NeurIPS 2023	Pairwise preferences on 10,000 instructions	$3,000 one time, at a median $21 an hour	About $0.30 a comparison (derived)

The AlpacaFarm team recruited annotators from Mechanical Turk with a 25-question qualification test and kept 16 of 34 whose agreement with the authors exceeded 70% [18]. Screening is part of the price.

Experts cost more. The GPQA authors (November 2023) hired 61 contractors through Upwork "to write and validate the dataset" and estimate "an average hourly payment of approximately $95 per hour" across that work, with a maximum possible rate of $150 an hour [19]. Writers received a $10 base payment per question plus bonuses tied to validators' answers, and the team collected 564 questions [19]. Humanity's Last Exam (January 2025) set up "a $500,000 USD prize pool, with prizes of $5,000 USD for each of the top 50 questions and $500 USD for each of the next 500 questions" for a benchmark of 2,500 questions [20]. Both are evaluation sets.

Checking the labels. Veselovsky et al. (June 2023) reran a summarization task on Mechanical Turk and estimated that "33-46% of crowd workers used LLMs when completing the task" [21]. Northcutt et al. (NeurIPS 2021) estimated "an average of at least 3.3% errors across the 10 datasets" whose test sets they studied, with label errors making up at least 6% of the ImageNet validation set [22]. Paying for human labels now includes checking that they are human and correct.

The lost baseline. This report infers that cost comparisons which priced human labels on Mechanical Turk lost their live reference price on 30 September 2026 [14]. Two of the model-label studies reviewed here compared against Mechanical Turk workers [18, 23]. No source reviewed measures label prices after the closure.

5. Model-made labels and their measured cost

The same studies priced labels from models, at the model prices of their year.

Study	Published	Task	Model label cost	Human comparison
Wang et al. [24]	August 2021	SST-2 and other language tasks	"as low as 0.002 dollars on average" a label on SST-2	$0.11 a label (estimated), 50% to 96% total saving at equal downstream performance
Ding et al. [25]	June 2023	SST-2 sentiment	$11.31 for 3,000 labels, $22.63 for 6,000	$221 to $300 for 3,000 human labels (estimated)
Gilardi et al. [23]	July 2023, PNAS	Tweets and news articles, 6,183 documents	Less than $0.003 per annotation	"about thirty times cheaper than MTurk"
Dubois et al., AlpacaFarm [18]	NeurIPS 2023	Pairwise preferences	$6 for 1,000 outputs	"50x cheaper than human annotation"

Gilardi et al. used a commercial chat model family. Its zero-shot accuracy exceeded that of crowd workers "by about 25 percentage points on average", and intercoder agreement was "about 56% for MTurk, 79% for trained annotators" [23].

Cheap labels do not always match the full human set. In Ding et al., 6,000 model labels reached 89.29 accuracy against 88.47 for 3,000 human labels, "approximately 10% of the cost of human labeling" [25]. The full set of 67,349 human labels reached 93.52 [25]. Model labels cost about $0.004 each in that study (derived) [25]. Wang et al. reported savings of 96% on SST-2, 93.8% on Gigaword summarization and 50% to 75% on other tasks, counting only "the labeling cost charged per label" and leaving out template and labeler selection [24].

Generated data. Stanford's Alpaca team, in March 2023, generated "52K unique instructions and the corresponding outputs" from 175 human-written seed pairs for "less than $500" through a commercial model's API, and fine-tuned on them in 3 hours on 8 80GB A100s for "less than $100 on most cloud compute providers" [26]. The team restricted the result to academic research and prohibited commercial use, in part because the data came from a model whose terms of use "prohibit developing models that compete" with its maker [26]. Generated data can be cheap in cash and costly in rights.

The limit. Shumailov et al. found in a 2024 Nature paper that "indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear" [27]. In July 2024 they expected data about genuine human interactions to grow in value as model-generated content spreads through crawled data [27].

6. What buyers and investors paid labeling companies

A public data services vendor. Innodata's 10-K for 2025, filed 26 February 2026, reports total revenue of $251.7 million in 2025 against $170.5 million in 2024, an increase of about 48%, with its Digital Data Solutions segment at $220.9 million against $141.1 million [29]. Customers engage it for "training and post-training data development" and related work [29]. One segment customer accounted for about 58% of total revenue in 2025, up from 48% in 2024 [29]. Its agreements are master service agreements with project statements of work that "generally do not obligate customers to purchase services in future periods" and are "generally terminable by our customer upon 30 to 90 days' notice" [29]. The figures show what buyers paid one vendor in total. They give no price per label or per project, and the filing names no buyer.

A valuation. Scale AI announced on 12 June 2025 "a significant new investment" that "values Scale at over $29 billion", with the investor holding "a minority of Scale's outstanding equity" [28]. A valuation is no price for data and appears here as context only.

7. Upkeep, the cost after delivery

Licenses run out. Reddit's remaining performance obligations in contracts longer than a year stood at $143.7 million at 31 December 2025, "primarily of long-term content licensing contracts", of which Reddit expected to recognize $118.9 million in 2026 and $24.8 million in 2027 [2]. By 30 September 2026 the figure was $62.4 million [3]. The 10-Q states that "our agreements with those partners are approaching the end of their respective current terms" [3]. For a buyer, a term license is a cost that returns at renewal, on terms the filing says may differ [2].

Delivery costs money. CuriosityStream reports that other cost of revenues rose with "higher hosting and web service costs directly related to the delivery and management of data assets" [6]. Wiley recognizes some licenses at a point in time [5], while Reddit delivers over the term [1], so the structure decides whether a cost is paid once or carried.

Evaluation recurs. AlpacaFarm reports "a recurring $242 cost for evaluating a single model on the 805 evaluation instructions" with human annotators, about $0.30 an instruction (derived) [18]. Every new model version repeats it.

Errors need rework. Northcutt et al. had crowd workers validate algorithmically flagged label errors and confirmed 51% of the candidates on average [22]. Finding and fixing errors is labor paid after the first delivery.

Maintenance is a standing cost. Sculley et al. wrote in 2015 that "it is common to incur massive ongoing maintenance costs in real-world ML systems", naming "data dependencies, configuration issues, changes in the external world" among the risk factors [30]. Sambasivan et al. (CHI 2021) interviewed 53 practitioners in India, East and West African countries and the USA and reported data cascades, "compounding events causing negative, downstream effects from data issues", at 92% prevalence [31]. Datasheets for Datasets

(revised December 2021) asks "dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers", including who hosts the dataset and how often it is updated to correct labeling errors [32].

8. Supply limits on public text

Villalobos et al. projected in June 2024 that, if current LLM development trends continue, models are trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained [33]. "The median exhaustion year is 2028, and by 2032 exhaustion becomes very likely", against an effective stock of indexed web text of around 4e14 tokens, about 400 trillion [33].

Access to public text is narrowing too. Longpre et al., in a July 2024 paper, audited 14,000 web domains and found that "in a single year (2023-2024) there has been a rapid crescendo of data restrictions from web sources, rendering ~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use" [34]. "For Terms of Service crawling restrictions, a full 45% of C4 is now restricted" [34].

Supply grows at the same time, by 3 to 5 billion pages a month at Common Crawl [10]. Read together, these sources describe public text that is restricted more often and human data that its holders license for a price [7, 27, 34].

Implications

For product leads and founders budgeting a dataset. The cost of a dataset follows its scope. The sources decide what licensing costs and whether a license is a fixed sum, a term contract or a revenue share [1, 5, 7]. The fields and labels decide labeling cost, which in the studies reviewed ranged from fractions of a cent per model label to about $95 an hour for expert question writers and validators [19, 23, 25]. Free sources add compute for filtering and scoring, and a check of their stated licenses [12, 35]. History and refresh decide how often content is bought, labeled and delivered again [2, 3, 18]. A custom dataset from Rillor's data services is specified by the same items: the sources in scope, the entities and fields, the format, how far back the records reach, the regions, the refresh cadence and the delivery.

For finance staff reading price evidence. A contract value [1], a revenue line [2, 4], a settlement payment [9] and a valuation [28] are four different numbers, and none is a unit price for data.

For teams choosing between labels from people and from models. The measured savings from model labels are large and dated. They come from studies of 2021 to 2023 on text tasks, at the model prices of their year, and they leave out setup and selection costs [23, 24, 25]. In one study the full human set still scored higher than the model-labeled sets [25]. Human labels now need checks for model use and for errors [21, 22].

For anyone planning beyond one delivery. Renewals, refreshes and repeat evaluations are costs after delivery [3, 18, 30], and the June 2024 supply projection [33] and the measured web restrictions [34] bear on what later public data costs.

Limits of this analysis

- No filing reviewed discloses a price per record, token, image, hour or title for licensed content.
- The revenue lines of Reddit, Shutterstock and CuriosityStream include products other than training data [2, 4, 6].
- The per-label figures come from studies of 2021 to 2023 on text tasks [18, 23, 24, 25]. Current model prices per label were not gathered.
- No source reviewed measures label prices after Mechanical Turk closed. The lost-baseline point is this report's inference.
- The GPU hours reported for FineWeb carry no dollar figure here, because no price source for them was gathered [12].
- No primary source reviewed gives the total size of the training data market, and analyst estimates were not used.
- No public primary source reviewed gives the full cost of a commissioned dataset. This report shows its parts.
- The Copyright Office report is cited in its pre-publication version of 9 May 2025, the only one listed on 8 November 2026 [7, 36].
- This report is research. It gives no legal advice and no view on any company named.

Sources

1. Reddit, Inc. Registration Statement on Form S-1. U.S. Securities and Exchange Commission, 22 February 2024, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1713445/000162828024006294/reddits-1q423.htm

2. Reddit, Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, 6 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1713445/000171344526000022/rddt-20251231.htm

3. Reddit, Inc. Quarterly Report on Form 10-Q for the quarter ended 30 September 2026. U.S. Securities and Exchange Commission, 30 October 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1713445/000171344526000116/rddt-20260930.htm

4. Shutterstock, Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, 17 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1549346/000154934626000008/sstk-20251231.htm

5. John Wiley & Sons, Inc. Annual Report on Form 10-K for the year ended 30 April 2026. U.S. Securities and Exchange Commission, 24 June 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/107140/000162828026045111/jwa-20260430.htm

6. CuriosityStream Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, 12 March 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1776909/000162828026017346/curi-20251231.htm

7. Copyright and Artificial Intelligence, Part 3: Generative AI Training (pre-publication version). U.S. Copyright Office, 9 May 2025, viewed 8 November 2026. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

8. Anthropic Copyright Settlement. Settlement Administrator, viewed 8 November 2026. https://www.anthropiccopyrightsettlement.com/

9. Notice of $1.5 Billion Proposed Class Action Settlement, Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal.), long-form notice. Settlement Administrator, 18 November 2025, viewed 8 November 2026. https://assets-us-01.kc-usercontent.com:443/1eeb16db-4934-006e-40a6-38fa91285ebb/ce109a57-390c-4d76-ad65-5807d4830fe5/ANT%20-%20Long-Form%20Notice_Updated%20Deadline%201.8.2026.pdf

10. Common Crawl. Common Crawl Foundation, viewed 8 November 2026. https://commoncrawl.org/

11. Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. Stefan Baack and Mozilla Insights, Mozilla Foundation, 6 February 2024, viewed 8 November 2026. https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/common-crawl/

12. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Guilherme Penedo and others, NeurIPS 2024 Datasets and Benchmarks Track, arXiv 2406.17557, version 2, 31 October 2024, viewed 8 November 2026. https://arxiv.org/abs/2406.17557

13. The rising costs of training frontier AI models. Ben Cottier and others, arXiv 2405.21015, version 2, 7 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2405.21015

14. Amazon Mechanical Turk Help. Amazon Mechanical Turk, viewed 8 November 2026. https://www.mturk.com/help

15. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. Kotaro Hara and others, CHI 2018, arXiv 1712.05796, 14 December 2017, viewed 8 November 2026. https://arxiv.org/abs/1712.05796

16. Digital labour platforms and the future of work: Towards decent work in the online world. Janine Berg and others, International Labour Office, 20 September 2018, viewed 8 November 2026. https://www.ilo.org/publications/digital-labour-platforms-and-future-work-towards-decent-work-online-world

17. What is your pricing? Prolific Researcher Help Centre. Prolific, viewed 8 November 2026. https://researcher-help.prolific.com/en/articles/445239-what-is-your-pricing

18. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. Yann Dubois and others, NeurIPS 2023, arXiv 2305.14387, version 4, 8 January 2024, viewed 8 November 2026. https://arxiv.org/abs/2305.14387

19. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. David Rein and others, arXiv 2311.12022, 20 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2311.12022

20. Humanity's Last Exam. Long Phan and others, arXiv 2501.14249, 24 January 2025, latest version 19 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2501.14249

21. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. Veniamin Veselovsky, Manoel Horta Ribeiro and Robert West, arXiv 2306.07899, 13 June 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.07899

22. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Curtis G. Northcutt, Anish Athalye and Jonas Mueller, NeurIPS 2021 Datasets and Benchmarks, arXiv 2103.14749, version 4, 7 November 2021, viewed 8 November 2026. https://arxiv.org/abs/2103.14749

23. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli, Proceedings of the National Academy of Sciences 120(30), arXiv 2303.15056, version 2, 19 July 2023, viewed 8 November 2026. https://arxiv.org/pdf/2303.15056v2

24. Want To Reduce Labeling Cost? GPT-3 Can Help. Shuohang Wang and others, arXiv 2108.13487, 30 August 2021, viewed 8 November 2026. https://arxiv.org/abs/2108.13487

25. Is GPT-3 a Good Data Annotator? Bosheng Ding and others, ACL 2023, arXiv 2212.10450, version 2, 14 June 2023, viewed 8 November 2026. https://arxiv.org/abs/2212.10450

26. Alpaca: A Strong, Replicable Instruction-Following Model. Rohan Taori and others, Stanford Center for Research on Foundation Models, 13 March 2023, viewed 8 November 2026. https://crfm.stanford.edu/2023/03/13/alpaca.html

27. AI models collapse when trained on recursively generated data. Ilia Shumailov and others, Nature 631, 755 to 759, 24 July 2024, viewed 8 November 2026. https://www.nature.com/articles/s41586-024-07566-y

28. Scale AI Announces Next Phase of Company's Evolution. Scale AI, 12 June 2025, viewed 8 November 2026. https://scale.com/blog/scale-ai-announces-next-phase-of-company-evolution

29. Innodata Inc. Annual Report on Form 10-K for the year ended 31 December 2025. U.S. Securities and Exchange Commission, 26 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/903651/000110465926020655/inod-20251231x10k.htm

30. Hidden Technical Debt in Machine Learning Systems. D. Sculley and others, Advances in Neural Information Processing Systems 28 (NIPS 2015), NeurIPS, 2015, viewed 8 November 2026. https://papers.nips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

31. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Nithya Sambasivan and others, CHI 2021, ACM, 2021, viewed 8 November 2026. https://research.google/pubs/everyone-wants-to-do-the-model-work-not-the-data-work-data-cascades-in-high-stakes-ai/

32. Datasheets for Datasets. Timnit Gebru and others, arXiv 1803.09010, version 8, 1 December 2021, viewed 8 November 2026. https://arxiv.org/abs/1803.09010

33. Will we run out of data? Limits of LLM scaling based on human-generated data. Pablo Villalobos and others, arXiv 2211.04325, version 2, 4 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2211.04325

34. Consent in Crisis: The Rapid Decline of the AI Data Commons. Shayne Longpre and others, arXiv 2407.14933, 24 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.14933

35. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre and others, arXiv 2310.16787, 4 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2310.16787

36. Copyright and Artificial Intelligence. U.S. Copyright Office, viewed 8 November 2026. https://www.copyright.gov/ai/

About this report

Published by Rillor on 9 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 training data costs, from filings and studies. 9 November 2026. https://rillor.com/insights/what-training-data-costs

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