

What AI training data sells for: a sourced index of public prices, 2023 to 2026

Working paper; the same file is intended for RePEc via MPRA (target 4.14)

Yuma Heymans

AIData.Broker, <https://aidata.broker> | hello@aidata.broker | 18 October 2026

Abstract

Training data for machine learning is bought and sold without an exchange or a posted price list. This paper assembles the public record of its prices: 1,022 observations for nine kinds of data, read from 735 source documents between 11 and 17 October 2026, 1,009 of them dated 2023 or later. Each observation keeps the figure as published, the side of the market where the money moves (paid to makers, offered to holders, asking price, made to order, completed sale, cost to produce) and a conversion to a common unit by stated arithmetic, with ranges entered as geometric midpoints. Medians in each kind's main unit run from \$0.525 per image to \$316,228 per company archive, and within a main unit the upper quartile is 2.0 to 37.8 times the lower. The side of the market moves prices as much as the kind of data does. For agent-training tasks, the median price buyers pay for collection to order is 9.5 times the median paid to the people who make them. For company records, the median published offer to holders is 11.2 times the median of the four completed sales with a stated amount. Completed sales are 73 of the 1,022 observations, and only 11 observations come from court filings. We describe the disclosure, promotion, concentration and survivorship biases of a public-record compilation and state the method so that every figure can be recomputed.

JEL codes: D40, D83, L14, L86, O33

Keywords: training data; data markets; price index; data licensing; artificial intelligence; market transparency; posted prices

1. Introduction

What does a dataset sell for when the buyer intends to train a model on it? The question matters to holders, to buyers and to anyone who must value data in a sale, an audit or a dispute. It has no simple answer, because no venue posts and compares prices. Sales are bilateral, terms are confidential, and the goods are heterogeneous: an hour of speech, a book, a repository and a company's email archive are not substitutes, and two email archives rarely are.

The public record is thin but not empty. Securities filings disclose licensing revenue when it is material. Court dockets disclose auction results and settlements. Catalogues and marketplaces post asking prices. Vendors publish rate cards and annotated price guides. Platforms that recruit workers post hourly rates. Since 2026, buyers of company records publish payout ranges and online estimators. Each source is partial, and each speaks from one side of the market.

This paper describes a compilation of every public price observation its compiler could source, one row per published figure with the source's own words; states a normalisation to common units by arithmetic on the stated numbers, with a summary statistic suited to ranges that span orders of magnitude; and separates the sides of the market. The findings are descriptive. Five stand out.

1. Prices exist for nine kinds of data, each in its own natural unit, and the medians in those units span nearly six orders of magnitude, from \$0.525 per image to \$316,228 per company archive.
2. Within a kind and unit, the upper quartile is 2.0 to 37.8 times the lower quartile. The narrowest spreads are in hourly pay for human work; the widest are where the unit itself is heterogeneous (a repository, a title, an image).
3. The side of the market matters as much as the kind of data. Research-only licences for recorded speech list at a median of \$13.91 an hour against \$118 for commercial listings. For agent-training tasks, collection to order prices at 9.5 times what the makers are paid.
4. Published offers are not realised prices. For company records the median published offer is \$316,228 across 93 figures, while the four completed sales with a stated amount have a median of \$28,311, and the one fully public auction shows how non-price terms moved the result.
5. The record over-represents what sellers of services and buyers seeking supply choose to publish, and under-represents completed sales, which are 73 of 1,022 observations.

2. Data and method

2.1 The compilation

The data are the "AI training data prices" dataset published by AIData.Broker [1], reviewed on 17 October 2026 and read for this paper on 18 October 2026, with its companion table of market rates [2]. Both are released under CC BY 4.0. The author's company compiles them; see the disclosure at the end.

An observation is a figure somebody else published for data used, or offered for use, in training machine learning models. Each row records the figure as the source writes it and as a low and a high number; the unit as written ("per corpus of about 364 hours") and a normalised unit; the figure per normalised unit, with any arithmetic in words; the side of the market and the basis of the source; the date of the event or publication and the date the source was read; and the source's URL and its own words (1,010 of the 1,022 rows carry a verbatim quote).

The basis records the kind of document: a court record, the payer's own page, a press report, a vendor's published price guide, or a catalogue or marketplace listing. The side records where in the chain from data maker to model developer the money moves: **paid to makers** (a vendor or platform pays a worker, speaker or collector); **offered to holders** (a buyer's public offer to the owner of existing data, including estimator outputs); **asking price** (a catalogue or marketplace listing); **made to order** (a buyer pays a vendor to collect or make data to a specification); **completed sale** (a licence, winning bid or settlement, as disclosed); and **cost to produce** (making data with no seller, such as synthetic data or compute).

Observations replaced by a newer reading of the same figure are kept by the compiler as history and are not in the published table. Prices that buy research or non-commercial use only (64 rows, mostly academic catalogue fees) are flagged and kept apart from commercial prices in every statistic below.

This paper takes the rows as compiled. It re-derives every statistic from them, but it does not re-read each of the 735 sources.

Table 1 shows the composition of the record by kind of data and side.

Table 1. Observations by kind of data and side of the market

Kind of data	Paid to makers	Offered to holders	Asking price	Made to order	Completed sale	Cost to produce	All
Company records	0	103	0	0	8	0	111
Source code	12	12	8	2	2	1	37
Video and footage	9	9	28	11	11	0	68
Audio and speech	35	0	149	10	9	2	205
Images	11	6	39	8	10	0	74
Text and archives	3	10	67	2	30	1	113
Datasets, tasks and environments	61	0	0	25	1	12	99
Physical-world data	49	0	33	44	2	11	139
Work by people	122	1	0	53	0	0	176
All kinds	302	141	324	155	73	27	1,022

Source: [1]; counts of rows by the dataset's group and side fields.

Asking prices (324) and pay to makers (302) dominate, because their posters want to sell or recruit; completed sales are 73 rows, 7.1 percent. Sides are unevenly spread: company records are almost all published offers, audio mostly catalogue listings, human work mostly posted pay.

Table 2 cross-tabulates side and basis.

Table 2. Observations by side and basis of the source

Side	Court record	Payer's own page	Press report	Vendor price guide	Catalogue listing	All
Paid to makers	0	232	28	1	41	302
Offered to holders	5	85	24	27	0	141
Completed sale	6	22	43	2	0	73
Asking price	0	6	11	6	301	324
Made to order	0	17	30	99	9	155
Cost to produce	0	9	3	15	0	27
All	11	371	139	150	351	1,022

Source: [1]; counts of rows by side and basis.

Completed sales rest mostly on press reports (43 of 73). Only 11 observations in the whole record come from court filings, and all of them are offers or completed sales.

2.2 Normalising units

Sources quote prices in the unit their trade uses: per minute or per hour of media, per corpus, per image, per word, per title, per task, per company, per year of a contract. The compilation maps each row to one of 20 normalised units and records the figure per that unit. Where the stated figure is already in the normalised unit, it is carried as is; this is the case for 661 rows. Where it is not, the conversion uses only

numbers the source states: 313 rows needed one. For 48 rows the source does not give what the arithmetic needs, and no per-unit figure is recorded.

Four examples show the arithmetic. The Linguistic Data Consortium lists Call My Net 1, 364 hours of conversational telephone speech, at a \$2,500 non-member research fee [3]: $\$2,500 / 364 = \$6.87$ an hour. A video creator reported \$33,000 from licensing about 1,000 hours of old videos in a year [4]: \$33 an hour. The Bartz v. Anthropic settlement fund of \$1.5 billion covers 482,460 listed works [5]: \$3,109.07 a work before fees. A reported licence of "close to \$250 million over five years" becomes \$50 million a contract year [6].

Figures in euros were converted at the European Central Bank reference rate of 1.1206 US dollars per euro of 9 October 2026, as recorded in each row's note; other currencies carry their conversion in the row. Per-minute media prices are multiplied to the hour. No figure is adjusted for inflation.

Units are not all units of data. A contract, a contract year or a whole dataset is a unit of a deal, and its size varies with what the deal covers. These three units are excluded wherever this paper compares kinds of data on one scale.

2.3 One figure per observation

Many sources quote a range ("\$100K to \$5M per company"), a floor ("\$1,000,000+") or a ceiling ("earn up to \$1M"). Each closed range is summarised by its geometric midpoint, the square root of the product of its ends; a point is taken as stated; where the two ends are equal or the lower end is zero, the arithmetic mean is used. Prices in this market are quoted in ranges whose ends differ by an order of magnitude or more, and the geometric midpoint respects that. For a range of \$100,000 to \$5,000,000 the geometric midpoint is \$707,107, against an arithmetic midpoint of \$2,550,000 that would sit closer to the top of the range than to its centre in proportional terms.

Open-ended figures, with only a floor or only a ceiling, are counted and reported as open but are not averaged. There are 83 such rows. With the 48 rows that have no per-unit figure at all, 131 rows are left out of the averages, and 891 rows carry one figure per normalised unit.

2.4 Cells, statistics and spreads

A cell is a kind of data, a side, a normalised unit and a use (commercial or research-only). The compilation's 1,022 rows form 115 cells. Within a cell, the median and quartiles are computed by linear interpolation between order statistics at position $(n - 1)q$. A median is reported only where a cell holds three or more figures; 58 of the 115 cells qualify.

Each kind's main unit is the normalised unit holding most of its commercial figures, contracts, contract years and whole datasets aside; main-unit statistics pool all sides.

A spread between two sides is the ratio of their medians in the same kind and unit. It is reported only for like-for-like pairs: what makers are paid against what buyers pay for collection to order (data made to order), and what holders are offered against asking prices and completed sales (data that already exists). A spread is treated as robust only where both cells hold eight or more figures.

2.5 Replication

Every statistic here was recomputed with a short Python script from the published JSON [1]; the count and median of each of the 115 cells match the compiler's market-rates table [2] exactly.

3. Findings

3.1 What has a public price, and in what unit

Table 3 places the nine kinds of data on one scale, each in its main unit.

Table 3. Commercial public prices in each kind's main unit (all sides pooled)

Kind of data	Main unit	n	Median	Middle half (p25 to p75)	p75/p25	Lowest to highest	Orders of magnitude
Company records	company	97	\$316.2K	\$152.7K to \$692.8K	4.5	\$9,381 to \$12.50M	3.1
Source code	repository	9	\$1,579	\$837 to \$31.6K	37.8	\$98.70 to \$65.8K	2.8
Video and footage	hour of video	45	\$37.14	\$14.65 to \$101	6.9	\$2.60 to \$632	2.4
Audio and speech	hour of audio	132	\$118	\$63.75 to \$266	4.2	\$0.520 to \$3,609	3.8
Images	image	46	\$0.525	\$0.131 to \$3.15	24.0	\$0.007 to \$40.00	3.8
Text and archives	book or title	18	\$515	\$105 to \$2,875	27.4	\$6.92 to \$5,000	2.9
Datasets, tasks and environments	environment	33	\$2,000	\$1,500 to \$3,000	2.0	\$500 to \$77.5K	2.2
Physical-world data	hour of video	56	\$16.16	\$5.75 to \$39.29	6.8	\$0.581 to \$340	2.8
Work by people	hour of work	114	\$77.46	\$51.23 to \$101	2.0	\$1.62 to \$397	2.4

Source: computed from [1]. Commercial prices only; contracts, contract years and whole datasets excluded; one figure per observation (Section 2.3); figures above zero. "Orders of magnitude" is log10 of highest over lowest.

The range of medians reflects the size of the unit as much as the value of the data: a company archive bundles millions of messages and documents, while an image is one file. Each kind has settled on a unit its sellers and buyers both understand, and comparing kinds would require a model of how many units of one substitute for another, which the record does not supply.

The extremes in each row are informative about what the unit hides. The cheapest hour of audio, \$0.52, comes from a marketplace listing of 1,587,658 hours of Hindi call-centre recordings offered as one lot at \$826,700 [34]; the dearest, \$3,609, is the average price per hour of speech-synthesis datasets licensed in 2019 (274 hours), reported in a Chinese data vendor's audited IPO prospectus [35]. The cheapest image figure is the median payout per portfolio image, \$0.0069, that 58 contributors received from one stock agency's 2023 fund sharing its AI data-licensing revenue [36]; the dearest is \$40 per stained pathology

slide on a commercial licence [37]. The cheapest company figure, \$8,000 to \$11,000, is an estimator's output for a company of one to ten people that is winding down [16]; the highest is a \$12.5 million late proposal for a bankrupt airline's archive [7].

3.2 The spread within each kind

Three patterns appear in the p75/p25 column of Table 3.

First, prices for human work are the tightest: posted hourly pay has a middle half of \$51.23 to \$101, plausibly because the same people can move between platforms. The environment row is as tight (2.0) for another reason: 28 of its 29 pay observations come from one bounty programme (Section 4).

Second, prices for units that are not standard goods are the widest. A "repository" can be a student project or a production codebase; a "title" can be a trade paperback licensed by a publisher or a scholarly monograph whose author receives a share; an "image" can be a stock photograph or an annotated medical slide. Code (37.8), text (27.4) and images (24.0) are the three widest rows.

Third, segments within a cell explain part of the spread. Table 4 shows three segmentations.

Table 4. Medians by segment within three cells

Cell	Segment	n	Median
Audio, asking price per hour	Arabic	5	\$442
	Spanish	7	\$202
	English	15	\$192
	French	9	\$160
	Japanese	6	\$84.67
	Hindi	8	\$45.00
	Tamil	3	\$33.62
Human work, pay per hour	Finance	4	\$136
	Medicine	6	\$118
	Law	5	\$90.00
	Coding	4	\$81.16
	Red-teaming	9	\$69.28
	Languages	3	\$39.69
	Generalist	8	\$20.61
Company records, offers per company	20 employees	3	\$221.7K
	50 employees	3	\$415.7K
	100 employees	5	\$586.2K
	250 employees	3	\$1.24M

Source: computed from [1], commercial prices, segments as tagged in the dataset; segments with three or more figures. The company-records segments are mostly estimator outputs for stated inputs.

Each segmentation is small, but the ordering is plausible. Posted pay for specialists in finance and medicine is between five and seven times posted pay for generalists. Commercial speech listings in Arabic and Spanish sit above English, and listings in Hindi and Tamil sit far below; the language segment confounds scarcity with the wage level of the country where speech is recorded, which the record cannot separate. Offers for company records rise with headcount, which is by construction: the estimators that produce most of these rows take headcount as an input (Section 3.4).

3.3 The sides of the market

The same unit is priced at each point in the chain from maker to model developer. Table 5 sets the sides side by side for the eight units with the most observations.

Table 5. Medians by side within one unit (commercial prices)

Kind	Unit	Medians by side (n)
Audio and speech	hour of audio	Paid to makers \$30.00 (3); asking price \$118 (109); made to order \$122 (9); completed sale \$76.15 (9)
Video and footage	hour of video	Paid to makers \$2.97 (4); offered to holders \$94.87 (7); asking price \$37.14 (17); made to order \$31.62 (11); completed sale \$41.96 (6)
Physical-world data	hour of video	Paid to makers \$14.23 (14); asking price \$8.88 (16); made to order \$37.74 (20); cost to produce \$127 (4)
Images	image	Paid to makers \$0.536 (8); offered to holders \$0.349 (6); asking price \$1.71 (24); made to order \$0.439 (4); completed sale \$0.063 (4)
Text and archives	book or title	Offered to holders \$355 (3); asking price \$77.95 (4); completed sale \$2,750 (10)
Company records	company	Offered to holders \$316.2K (93); completed sale \$28.3K (4)
Datasets, tasks and environments	task	Paid to makers \$300 (11); made to order \$2,837 (8); cost to produce \$1,500 (5)
Work by people	hour of work	Paid to makers \$77.46 (105); made to order \$18.97 (9)

Source: computed from [1]. Cells with fewer than three figures are omitted.

Three readings follow.

Rights scope is a price. The record keeps research-only licences apart. For recorded speech, 36 research-only listings have a median of \$13.91 an hour (middle half \$6.86 to \$31.96), against \$118 for 109 commercial listings: commercial use lists at 8.5 times the research price in the same unit. The Linguistic Data Consortium, for example, licenses non-members for research use only and requires a for-profit membership for commercial use [3].

Making data to order costs more than paying the maker. For agent-training tasks, the median price buyers pay vendors for tasks made to order is \$2,837 (eight figures), against a median of \$300 paid to the people who make them (eleven figures): a ratio of 9.5, the largest robust spread in the record. For first-person and teleoperation video for robots, collection to order has a median of \$37.74 an hour against \$14.23 paid to the people who record it, a ratio of 2.7. Direct margin evidence agrees in direction. Appen, which defines gross margin as revenue less crowd expenses, reported 40.3 percent for 2025 [8], so its

clients paid about \$1.68 for each dollar its crowd earned. Handshake, in a job posting, described paying about \$60 million a month to more than 30,000 experts at a run rate of about \$1 billion [9]; on its own figures, about 72 percent of revenue passed to experts.

Some spreads are composition, not margin. For human work per hour, the made-to-order median (\$18.97) is a quarter of the pay median (\$77.46). That is not a negative margin. The pay cell is dominated by expert marketplaces recruiting specialists, with one platform contributing 26 of 111 rows, while the made-to-order cell holds annotation vendors' billing rates (from \$6 an annotator hour [10]), federal schedule ceiling rates for labelling roles and recruitment platforms' fees. The two cells price different labour. Only these three pairs (tasks 9.46, robot video 2.65, human work 0.24) have eight or more figures on both sides; the other like-for-like ratios that Table 5 allows rest on three to seven figures at one end.

Books show why the side definitions matter. Offers to holders of books are, in the record, what authors receive from publishers' licensing programmes: \$205 and \$355 per title in two rounds of one programme [11] and \$2,500 as the author's half of another [12]. Completed sales are the gross amounts: half of the ten figures are per-work amounts from the Bartz settlement (about \$3,000 per work [5]), and one is the same publisher's \$5,000 per title, split equally with authors who opt in [12]. The ratio of 7.75 between the two medians measures the gap between what an author receives through a publisher and what changes hands at the top of the chain.

3.4 Quoted and realised prices for company records

Company records, a business's operating archives (email, chat, tickets, documents, code, customer records), have public prices in the record only from 2026: 111 observations, with the published offers in the main unit dated 13 August to 17 October 2026. They show how quoted prices form in a new market.

Of 103 offers to holders, 61 are outputs of nine public estimators or calculators, each row the output for one set of stated inputs. Estimators are not independent observations. Two of them, at micro1 and Handshake, compute the same formula: a \$140,000 base, \$2,333 per employee, \$11,667 per year of records and \$15.75 per gigabyte; for 100 employees and ten years in the United States they display \$379,290 to \$888,088 and \$383,000 to \$897,000 respectively [13][14]. The same formula weights the United States at 1, Canada and Europe at 0.75 and other countries at 0.3 [13]. For one company profile (about 100 people, about ten years, United States), the low ends of four public estimators range from \$37,500 to \$800,000, a factor of about 21 [13][15][16]. One firm's own site configuration describes its referrer calculator as "a deliberately generous range a referrer can show an owner" [17].

Against these quotations, the realised record is sparse. Four completed sales carry a stated amount in the main unit: a winning bid of \$10 million for a bankrupt airline's archive [21], a reported \$10,000 to \$100,000 per company for nearly 100 sales by one wind-down service [18], the same service's average of about \$10,000 implied by about \$1 million recovered across those deals [19], and a founder's roughly \$25,000 for a codebase [20]. Their median is \$28,311, and the ratio of the offer median to it is 11.2. Two of the four rows describe the same set of deals, which Section 4 discusses. One buyer states that its smallest closed deal was \$20,000 while its headline offered "up to" a much larger sum [16].

The one fully public sale shows how terms other than price moved the result. Table 6 sets out the auction for the de-identified records of Spirit Airlines in its Chapter 11 case.

Table 6. The Spirit Airlines records auction, August to September 2026

Step	Bidder	Amount	Condition	Outcome
Opening bid	Google	\$5.0M	Cash; third-party de-identification paid by the buyer	Overbid
First overbid	Mercor	\$5.2M	Third-party de-identification on the opening bidder's form	Treated as the highest bid of the round
First overbid, alternative	Mercor	\$7.0M	De-identification with its own tools	Not selected
Later bid	Mercor	\$10.0M	De-identification with its own tools	Not selected
Winning bid	Google	\$10.0M	Delivery in de-identified form; buyer pays de-identification	Selected; court approval pending
Alternate bid	Mercor	\$7.5M	Backup if the winning sale does not close	Named alternate
Late proposal	micro1	\$12.5M	All cash, unsigned, filed after the auction	Not adopted as of the record

Sources: [21] (Dkt. 1463, Exhibits A and B), [22] (auction summary), [7].

The estate treated a \$5.2 million bid as higher than a \$7.0 million bid in the same round because of who would control de-identification; the auction summary quotes the debtors as finding that \$5.2 million "was the highest bid for that round despite the lower price" [22]. A competitive process took the opening bid from \$5.0 million to \$10.0 million in about two and a half hours [22]. A late proposal 25 percent above the winning bid was filed after the auction [7][22]. As of the record, the sale had not been approved; the hearing was moved from 14 October to 6 November 2026 [23].

3.5 What drives the differences

The record does not support a hedonic regression: the cells are small, the units differ, and the sources do not report characteristics consistently. What it supports is a list of price drivers that sources state explicitly, with figures where they give them.

- **Exclusivity.** Two founders of environment-building companies told Epoch AI that exclusive deals are "roughly 4-5x more expensive than non-exclusive ones" [24], the only published multiple in the record.
- **Control of de-identification,** priced in the Spirit auction [22].
- **Connected systems.** One buyer's estimator values a 51 to 100 person company with four or more years of records at \$85,000 to \$121,000 with six connected tools and \$16,000 to \$22,000 with three [16].
- **Substitutes.** One vendor offers off-the-shelf call-centre audio at \$5 to \$25 an hour against \$50 to \$300 for custom collection [25]; Epoch AI reports about \$20,000 per website replica built for agent training [24], which caps what the original site's owner can be paid.
- **Metadata.** A media data broker's chief executive said it can charge \$1 to \$1,000 per hour of video "depending on the application and the available tagging and metadata" [26].
- **Net volume.** A wind-down service found that "roughly 37% of raw code disappeared after removing third-party/generated code and internal duplication" [27].
- **Compute parity.** A task vendor argued that labs should not spend "significantly less than ~\$500 per RL task" given about \$480 of compute per training run [28]; Mercor's chief executive described a \$1,500 cost basis for a task of ten hours at \$150 an hour [29].
- **Rights scope,** as in the research and commercial speech prices above.

These drivers are mostly statements by sellers or intermediaries about their own pricing. They describe how prices are argued, which is not the same as how they are set.

3.6 Disclosed licences of content

Large licences of content appear mainly through filings and the press. A separate ledger of disclosed licensing deals [30] holds 111 deals from July 2023 to 15 October 2026. Only 19 carry a dollar total or an annual amount. Table 7 counts them by half-year.

Table 7. Disclosed AI data licensing deals and those with a dollar figure

Half-year	Deals disclosed	With a dollar total or annual amount
2023 H2	4	1
2024 H1	21	6
2024 H2	12	2
2025 H1	9	0
2025 H2	27	1
2026 H1	16	4
2026 H2 (to 15 October)	22	5
All	111	19

Source: computed from [30], by the half-year of each deal's date.

The priced deals are large and few. Reddit's prospectus disclosed \$203.0 million of aggregate contract value in data licensing agreements with terms of two to three years [31]. News Corp's licence to OpenAI "could be" close to \$250 million over five years, as reported from The Wall Street Journal [6]. Reddit's licence to Google was reported at about \$60 million a year [32]. Shutterstock earned \$104 million of AI licensing revenue in 2023 [33]. In the price compilation, nine annual licence values for text have a median of \$16.0 million a year. Of the 111 deals, 40 license content for display or grounding in AI answers rather than for training, which is a different product with a different price; 43 are for training, 13 for both and 15 do not say.

4. Limitations: the biases of a public-record compilation

Disclosure bias. Prices enter the record when someone has a reason to publish them: sellers post asking prices, platforms post pay to recruit, buyers of company records post offers to attract supply. Completed sales are disclosed when a filing requires it (securities filings, bankruptcy dockets, class settlements) or a party speaks to the press. Completed sales are 73 of 1,022 rows, and 43 of those rest on press reports, some citing people familiar with the deals rather than named sources. The record therefore describes the posted market better than the transacted one.

Promotional bias. Many offers are marketing. Headline ceilings ("earn up to") and estimator outputs are designed to attract holders, and one firm states that its referrer calculator is deliberately generous [17]. Self-reported figures, such as a vendor's own payout totals or case studies with no named counterparty, are labelled as such in the record but are counted like other observations.

Source concentration. A cell's median can rest on one publisher. Table 8 shows the share of the largest single source host in eight main cells.

Table 8. Concentration of sources in eight cells

Kind	Side	Unit	Rows	Distinct source hosts	Rows from the largest host
Company records	Offered to holders	company	101	26	12 (12%)
Audio and speech	Asking price	hour of audio	110	8	55 (50%)
Work by people	Paid to makers	hour of work	111	22	26 (23%)
Datasets, tasks and environments	Paid to makers	environment	29	2	28 (97%)
Video and footage	Asking price	hour of video	22	7	14 (64%)
Physical-world data	Made to order	hour of video	20	5	10 (50%)
Images	Asking price	image	33	6	13 (39%)
Text and archives	Completed sale	book or title	12	8	4 (33%)

Source: computed from [1]; rows include open-ended figures; hosts are the source URLs' domains.

Half of the commercial speech listings come from one European catalogue, and nearly all environment pay figures come from one public bounty sheet. Those cells describe a publisher's price list, not a market.

Non-independence. Several rows describe the same underlying transactions. The wind-down service's nearly 100 deals appear once as a quoted range and once as an implied average. One publisher's per-title fee appears as a completed sale and its author's half as an offer. Five of the ten completed sales per title are figures from one settlement. The estimator rows are many outputs of a few formulas, two of them identical. A median over such rows weights a formula by the number of input profiles the compiler chose to run through it.

Survivorship. The record holds what was online between 11 and 17 October 2026. Nine postings found closed are kept at their last reading; pages withdrawn earlier are absent, and offers that produced no sales are not distinguished from offers that did.

Time. Of 1,022 observations, 848 are dated 2026 and 561 are dated October 2026, because listings and posted pay carry the day they were read rather than the day the price was first set. The dated record is therefore not an event-time series, and this paper does not estimate price trends. Thirteen observations predate 2023; the earliest is dated July 2017.

Normalisation. The geometric midpoint is a convention, and it misstates a row whose two ends describe different products: "about \$20K; about \$300K for a Slack-grade clone" enters as \$77,460 per environment. Conversions inherit any error in a source's own count of hours, works or files, and no adjustment is made for quality, licence length or exclusivity, which most sources do not state.

Small cells. Fifty-seven of the 115 cells hold too few figures for a median, and only three spreads have eight or more figures on both sides. The segment medians in Table 4 rest on three to fifteen figures each.

The compiler's position. The author leads a company that acts as a seller's agent in this market and compiles the datasets used here. That interest could bias which sources are sought and how rows are classified. The datasets are public, every row links to its source, and the method above is stated so that others can recompute or reclassify.

5. Conclusion

There is no price list for AI training data, but there is a public record of prices, and it can be organised. Organised by kind, side and unit, it shows a market in which each kind of data has its own unit, prices within a unit spread by factors of two to nearly forty, and the side of the market changes the price as much as the kind of data. Research and commercial rights are priced apart. Data made to order costs buyers a multiple of what its makers receive. Published offers for company archives sit an order of magnitude above the few completed sales with a stated amount, and the one public auction turned on who would control de-identification as well as on price.

The record's main weakness is the one that makes it necessary: completed sales are rarely disclosed. Three kinds of new disclosure would improve any index of this market: realised prices with their units and rights scope from intermediaries that run many small sales, court records of further asset sales in insolvency, and segment reporting of licensing revenue by listed rights holders. Until those exist, a public-record index is best read as a map of posted and quoted prices, with completed sales as sparse anchors, and every figure in it should be read with its side, its unit and its source.

Disclosure

The author is chief executive of AIData.Broker, which acts as a seller's agent for organisations that license data for model training and which compiles the datasets analysed here.

Data availability

The datasets [1][2][30] are published as JSON under CC BY 4.0, each row with its source URL, dates and the source's own words; the statistics here use the versions read on 18 October 2026. Browsable views: <https://aidata.broker/prices> and <https://aidata.broker/rates>; narrative guides on the same records: <https://aidata.broker/guides/ai-training-data-prices> and <https://aidata.broker/guides/how-much-is-company-data-worth>.

References

- [1] AIData.Broker, "AI training data prices" (dataset, 1,022 rows, reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/prices.json> (accessed 18 October 2026).
- [2] AIData.Broker, "AI training data market rates" (dataset, 115 cells, reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/market-rates.json> (accessed 18 October 2026).

- [3] Linguistic Data Consortium, "Call My Net 1" (LDC2024S05), catalogue entry with fees. <https://catalog.ldc.upenn.edu/LDC2024S05> (accessed 18 October 2026).
- [4] Semafor, "The AI boom is turning old content into cash cows for creators," 30 October 2025. <https://www.semafor.com/article/10/30/2025/the-ai-boom-is-turning-old-content-into-cash-cows-for-creators> (accessed 18 October 2026).
- [5] Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal.), Order Granting Final Approval of Class Action Settlement, Dkt. 680, 20 July 2026. https://assets-us-01.kc-usercontent.com:443/1eeb16db-4934-006e-40a6-38fa91285ebb/36cce252-a32c-4cbb-a624-31dd8ac9574c/2026-07-20%20Order%20Granting%20Final%20_dckt%20680_0_.pdf (accessed 18 October 2026); and Authors Guild, "Court grants final approval of Anthropic copyright settlement." <https://authorsguild.org/news/court-grants-final-approval-anthropic-copyright-settlement/> (accessed 18 October 2026).
- [6] VentureBeat, "OpenAI partners with Wall Street Journal publisher News Corp," 22 May 2024 (reporting The Wall Street Journal). <https://venturebeat.com/ai/openai-partners-with-wall-street-journal-publisher-news-corp> (accessed 18 October 2026).
- [7] Fortune, "micro1 tries to beat Google's bid to buy bankrupt Spirit Airlines data for \$12.5 million," 14 September 2026. <https://fortune.com/2026/09/14/micro1-tries-to-beat-googles-bid-to-buy-bankrupt-spirit-airlines-data-for-12-5-million-highlighting-value-of-data-to-ai-companies/> (accessed 18 October 2026).
- [8] Appen Limited, ASX announcement, FY25 results and FY26 outlook, 25 February 2026. <https://announcements.asx.com.au/asxpdf/20260225/pdf/06wq7ntr51bhgl.pdf> (accessed 18 October 2026).
- [9] Handshake, Engagement Manager job posting, 21 November 2025. <https://jobs.ashbyhq.com/handshake/f2ed7bba-e887-474d-9bf2-894a47f0082f> (accessed 18 October 2026).
- [10] Label Your Data, "Pricing." <https://labeyourdata.com/pricing> (accessed 18 October 2026).
- [11] Slate, "Arcadia Publishing: your town's local history books might be used to train A.I.," 3 November 2025. <https://slate.com/technology/2025/11/arcadia-publishing-artificial-intelligence-history-books-authors.html> (accessed 18 October 2026); and Cup of Coffee newsletter, 3 November 2025, reprinting Arcadia Publishing's email to authors. <https://www.cupofcoffeenews.com/cup-of-coffee-november-3-2025/> (accessed 18 October 2026).
- [12] Publishers Weekly, "Agents, authors question HarperCollins AI deal," 25 November 2024. <https://www.publishersweekly.com/pw/print/20241125/96533-agents-authors-question-harpercollins-ai-deal.html> (accessed 18 October 2026).
- [13] micro1, data payout estimator. <https://www.micro1.ai/data-payout> (accessed 18 October 2026; outputs recorded 11 October 2026).
- [14] Handshake AI, data partnerships page and estimator. <https://joinhandshake.com/ai/data-partnerships/> (accessed 18 October 2026).
- [15] Grepped, "Grepped for Business: payout estimate." <https://www.grepped.ai/business/> (accessed 18 October 2026); and Troveo, business data estimator. <https://www.troveo.ai/business-data> (accessed 18

October 2026).

[16] Oak Data, partner page and estimator. <https://www.oakdata.com/partner> (accessed 18 October 2026).

[17] Grepped, site configuration file. <https://www.grepped.ai/config.js> (accessed 18 October 2026; read 11 October 2026).

[18] Forbes, "AI's new training data: your old work Slacks and emails" (Anna Tong), 16 April 2026, <https://www.forbes.com/sites/annatong/2026/04/16/ais-new-training-data-your-old-work-slacks-and-emails/> (accessed 18 October 2026; figure as relayed in [7]).

[19] Upstarts Media, "Exclusive: SimpleClosure data on what AI labs buy," 3 September 2026. <https://www.upstartsmedia.com/p/exclusive-simpleclosure-data-what-ai-labs-buy> (accessed 18 October 2026).

[20] MIT Technology Review, "AI labs are buying data from dead startups to train their models," 13 October 2026. <https://www.technologyreview.com/2026/10/13/1145755/ai-labs-buying-data-from-dead-startups-to-train-models/> (accessed 18 October 2026).

[21] In re Spirit Aviation Holdings, Inc., No. 25-11897 (Bankr. S.D.N.Y.), Dkt. 1463, Notice of Auction Results, Exhibit A (Sale Agreement, Google LLC) and Exhibit B (Alternate Sale Agreement, Mercor.io). <https://storage.courtlistener.com/recap/gov.uscourts.nysb.331039/gov.uscourts.nysb.331039.1463.0.pdf> (accessed 18 October 2026).

[22] Chapter11Cases (Stretto), "The Spirit Airlines deidentified data sale: auction results, objections and the September 30 hearing," 13 September 2026. <https://chapter11cases.com/blogs/news/the-spirit-airlines-deidentified-data-sale-auction-results-objections-and-the-september-30-hearing> (accessed 18 October 2026).

[23] TravelPulse, "Spirit Airlines-Google \$10 million data sale delayed as lawmakers raise concerns." <https://www.travelpulse.com/news/airlines-airports/spirit-airlines-google-10-million-data-sale-delayed-as-lawmakers-raise-concerns> (accessed 18 October 2026).

[24] Epoch AI, "The state of RL environments," 12 January 2026. <https://epoch.ai/gradient-updates/state-of-rl-envs> (accessed 18 October 2026).

[25] AIxBlock, "Off-the-shelf call center audio vs custom." <https://aixblock.io/blogs/off-the-shelf-call-center-audio-vs-custom> (accessed 18 October 2026).

[26] The Media Leader, "TV rights owners promised new revenue stream if they license content for AI training," 29 January 2026. <https://uk.themedialeader.com/tv-rights-owners-promised-new-revenue-stream-if-they-license-content-for-ai-training/> (accessed 18 October 2026).

[27] SimpleClosure, "What your assets are worth," 3 September 2026. <https://simpleclosure.com/blog/insights/what-your-assets-are-worth/> (accessed 18 October 2026).

[28] Mechanize, "Cheap RL tasks will waste compute," 18 August 2025. <https://www.mechanize.work/blog/cheap-rl-tasks-will-waste-compute/> (accessed 18 October 2026).

- [29] Sequoia Capital, podcast with Brendan Foody (Mercor), "How RL environments are built and why they're your AI moat." <https://sequoiacap.com/podcast/how-rl-environments-are-built-and-why-they-re-your-ai-moat> (accessed 18 October 2026).
- [30] AIData.Broker, "AI data licensing deals" (dataset, 111 deals, reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/deals.json> (accessed 18 October 2026).
- [31] Reddit, Inc., Prospectus (SEC Form 424B4), March 2024. <https://www.sec.gov/Archives/edgar/data/1713445/000162828024012380/reddit-final424b4.htm> (accessed 18 October 2026).
- [32] The Verge, "Google and Reddit's AI training data deal," 22 February 2024. <https://www.theverge.com/2024/2/22/24080165/google-reddit-ai-training-data> (accessed 18 October 2026).
- [33] Bloomberg Law, "Shutterstock's AI licensing business generated \$104 million." <https://news.bloomberglaw.com/ip-law/shutterstocks-ai-licensing-business-generated-104-million> (accessed 18 October 2026).
- [34] Opendatabay, "1.5Mn+ Hours Hindi Dual-Channel Call Center Dataset" (Infobay.AI listing). <https://www.opendatabay.com/data/ai-ml/7cabed2a-cb16-4821-a02a-fd9426eff6fd> (accessed 18 October 2026).
- [35] Beijing Haitian Ruisheng Science Technology Ltd., STAR Market IPO prospectus (招股说明书), 9 August 2021, training data products table. <https://static.cninfo.com.cn/finalpage/2021-08-09/1210701086.PDF> (accessed 18 October 2026).
- [36] R. Kneschke, "Analyse: Wie viel zahlt Shutterstock für KI-Trainingsdaten an Anbieter?," Alltag eines Fotoproduzenten, 8 May 2023. <https://www.alltageinesfotoproduzenten.de/2023/05/08/analyse-wie-viel-zahlt-shutterstock-fuer-ki-trainingsdaten-an-anbieter/> (accessed 18 October 2026).
- [37] HistAI, DataHub whole-slide image pricing table. <https://www.hist.ai/> (accessed 18 October 2026).

Statements

Use of AI. This paper was prepared with generative AI tools, which compiled and analysed the open datasets and drafted the text. The author takes responsibility for its content.

Competing interests. The author is CEO of AIData.Broker, which acts for companies that license their data to AI developers and publishes the open datasets this paper draws on.

Licence: CC BY 4.0. Cite as: Heymans, Y. (2026). What AI training data sells for: a sourced index of public prices, 2023 to 2026. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-prices>