

How AI developers acquire training data: six channels and the shift away from the open web

Yuma Heymans

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

Abstract

Developers of large AI models obtain training data through six channels: collection from the open web without a licence, licences from publishers and platforms, commissioned human and expert work, the purchase or licence of organisations' internal records, purpose-built environments and synthetic data, and data from their own users and devices. This paper describes each channel from a public-record compilation read on 18 October 2026: 122 quantified supply facts, 111 disclosed licensing deals, 83 lawsuits, rulings and regulatory actions, 227 financial records of data vendors, 111 funding events, 110 brokerage records and 1,022 price observations, with a robots.txt census of 6,094 websites. The open web remains the largest source by volume, but access is narrowing: 11.4 percent of the census sites shut at least one AI crawler out of the whole site, and an earlier audit found restrictions on 28 percent or more of the most actively maintained sources of a standard web corpus. Of the 83 legal records, 70 concern content taken without a licence and 12 concern developers' use of their own users' data; none concerns the channels in which developers pay for data. Money is moving to those paid channels: disclosed licences, data vendors whose newest reported revenues or run rates exceed one billion dollars, and, since 2026, published programmes that buy organisations' records. We set out the scale, prices and legal exposure of each channel (not legal advice), the limits of the record, and what the shift changes for organisations that hold data.

Keywords: artificial intelligence; training data; data governance; web crawling; robots.txt; data licensing; platform labour; privacy; copyright; synthetic data

1. Introduction

Accounts of where AI training data comes from often begin and end with the web crawl. That account was adequate for the first generation of large language models, which were pretrained on cleaned copies of public web archives. It is no longer adequate. Developers now combine at least six channels, which differ in who must consent, who is paid, which body of law governs and how the data is produced.

1. **Collection without a licence:** crawling and scraping the open web and bulk copying of other public or unlicensed content.
2. **Licensing from publishers and platforms:** contracts that pay for access to archives, feeds or both.
3. **Commissioned human and expert work:** answers, rankings, reviews and tasks made to order by paid contributors, usually through vendors.
4. **Purchase of organisations' records:** licences or sales of the email, chat, tickets, documents and code that businesses accumulate in operation.
5. **Built environments and synthetic data:** simulated software in which models practise, and data generated by models.
6. **Users' own data:** what people post, write or record on products the developer or its affiliates operate.

This paper argues that the record of 2023 to 2026 shows a shift of spending, though not yet of volume, from the first channel to the second through fifth, under three pressures. The first is refusal: publishers and platforms increasingly tell crawlers not to collect, and some sue those that do. The second is the pricing of unlicensed use through litigation, of which the \$1.5 billion Bartz v. Anthropic settlement is the clearest figure. The third is a change in what models are trained to do. Post-training for reasoning, coding and agentic work needs data that a crawler cannot reach: expert answers, private records of how work is done, and environments with automatic graders. The sixth channel grows alongside, under data protection regulators' supervision rather than copyright litigation.

The contribution is empirical and organising. Each channel is described by its scale, its prices and its legal exposure, each quantity taken from a dated public source. The legal records are classified by channel, which shows where disputes concentrate. The paper closes with what the shift changes for organisations that hold data. It is not legal advice: where rules and cases are discussed, the aim is to describe the record, not to state how any law applies to any party.

2. Data and method

2.1 Sources

The paper draws on eight public datasets compiled by AIData.Broker and published under CC BY 4.0, each row carrying its source URL, dates and the source's own words [1], and on the compiler's daily robots.txt census [2]. All were read on 18 October 2026. Table 1 summarises them.

Table 1. Sources used

Source	Rows	Span of dates	What a row records
Supply facts	122	to 16 Oct 2026	A quantified fact about where a kind of data sits and how much exists
Licensing deals	111	Jul 2023 to 15 Oct 2026	A disclosed licence of content to an AI developer, its use and any value
Legal records	83	Sep 2023 to 15 Oct 2026	A lawsuit, ruling, settlement, regulatory action or guidance over training data
Vendor financials	227	FY2019 to Oct 2026	Revenue, run rate or other figure of a data or expert-work vendor
Funding events	111	Jan 2022 to 14 Oct 2026	A funding round, acquisition or investment in an AI data company
Brokerage log	110	Oct 2025 to 16 Oct 2026	A buyer programme, transaction, price range or legal development in company-data sales
Buyer board	37	read Oct 2026	A buyer or intermediary with a published way to acquire data
Price observations	1,022	2017 to 17 Oct 2026	A published price for training data, normalised per unit
Crawler census	3 sweeps	15 to 17 Oct 2026	robots.txt files read, and the sites that shut each AI crawler out

Sources: [1][2].

Each figure in this paper is cited to the original document the row names (a filing, a docket, a company page, a press report), with the dataset as the compilation. Press figures attributed to unnamed people are marked as such. Every count and share was computed with a short Python script from the published JSON.

2.2 Classifying legal records by channel

Each of the 83 legal records was assigned to one channel by the author, with a stated rule. Records about a developer's use of its own users' posts, conversations or device recordings were assigned to channel 6. One antitrust suit about AI summaries shown in search, which does not concern training, was set aside. All other records concern content collected without a licence and were assigned to channel 1; this includes privacy actions over scraped images of faces and contract suits by platforms against scrapers. No record concerned channels 2 to 5. The classification is a judgment; the identifiers of the records in each group are available from the author.

2.3 The crawler census

The census reads the robots.txt file of sites on the Tranco ranking of most visited sites (the copy of 15 October 2026), within its top 300,000, keeping addresses ending in .com, .net, .org, .io, .co, .ai, .app, .dev, .info, .us or .uk. It reads 4,000 sites each morning in list order, so that a full pass takes 46 mornings. A site counts as shutting a crawler out when its robots.txt names that crawler and disallows the whole site; a site with no robots.txt, or one that did not answer, is not counted [2]. The current pass had read three slices (the 42nd to 44th) on 15 to 17 October 2026: 12,000 files requested and 6,094 answered. Because the list is read in rank order, these slices lie in the lower part of the ranked list, and the shares below describe those sites rather than the most visited ones.

2.4 Measures that are not added up

Vendor revenues are reported in incompatible measures: audited annual revenue, gross revenue run rates, annual recurring revenue and self-reported annualised figures. This paper reports each with its measure and basis and does not sum them. Gross figures for expert marketplaces include what passes through to contributors.

3. The six channels

3.1 Collection without a licence

Scale. The open web is still the largest source by volume. Common Crawl's September 2026 archive holds 2.17 billion web pages (361.4 TiB uncompressed) from 40.4 million hosts [3]. Epoch AI estimated the effective stock of public human-generated text at around 300 trillion tokens and expected models to use it fully between 2026 and 2032 if trends continued; that estimate dates from June 2024 [4].

Diminishing novelty. Only 587.2 million of the September archive's URLs had not been crawled before [3], about 27 percent of the pages. Some sources of new public text are shrinking outright. Stack Overflow's own API counts 986 questions asked in September 2026 and still on the site, against 24,732 for September 2024 and 119,730 for September 2021 on the same query [5]: 4.0 and 0.8 percent of those months.

Refusal. A site states what crawlers may fetch in its robots.txt file. An audit of the C4 corpus found that restrictions added between 2023 and 2024 cut off about 5 percent or more of all its tokens and 28 percent or more of its most actively maintained, critical sources, while 45 percent of C4 was restricted by terms of service [6]. Table 2 reports the census.

Table 2. Share of sites that shut each AI crawler out of the whole site (crawler census, current pass)

Crawler	Operator and purpose	Sites	Share of 6,094
Any named AI crawler		696	11.4%
Two or more, at least one a training crawler		514	8.4%
GPTBot	OpenAI, training	457	7.5%
CCBot	Common Crawl, open archive	437	7.2%
Bytespider	ByteDance	427	7.0%
ClaudeBot	Anthropic	367	6.0%
meta-externalagent	Meta	327	5.4%
Google-Extended	Google, switch for model training	302	5.0%
Applebot-Extended	Apple, switch for model training	270	4.4%
PerplexityBot	Perplexity, answer engine	186	3.1%

Source: [2]; counts summed over the three sweeps of the current pass (Section 2.3).

Two features stand out. Refusal is a minority practice among these sites: nearly nine in ten of those that answered let every named AI crawler in. And sites refuse training crawlers more often than the crawler of an answer engine that cites and links its sources: GPTBot is refused by 7.5 percent, PerplexityBot by 3.1 percent. The economics of being crawled help explain refusal. Cloudflare measured Anthropic's crawlers fetching up to 500,000 pages for each visitor they referred back in 2025, settling at about 25,000 to 100,000 after May, while OpenAI's ratio peaked at 3,700 to one in March [7].

A robots.txt line is a request. A US court dismissed a claim that ignoring it circumvents an access control, holding that robots.txt directives do not effectively control access to copyrighted works [8]. Publishers and platforms have turned to copyright, contract and anti-scraping claims instead.

Cost and legal exposure. The channel has no licence fee; its cost is legal exposure, and the record now prices it. Of the 83 legal records, 70 concern this channel (Table 3). In *Bartz v. Anthropic*, the court held training on the plaintiffs' books to be fair use, and scanning purchased print books as well, but sent the pirated library copies to trial [9]; the case settled for \$1.5 billion, about \$3,000 per work, with final approval on 20 July 2026 [10]. Claims were filed for 440,490 of the 482,460 listed works, a 91.3 percent claims rate [11]. In *Kadrey v. Meta* the court granted Meta partial summary judgment on fair use on its record, covering 13 named authors [12]. The Third Circuit affirmed that a competitor's use of Westlaw headnotes to train a non-generative legal research tool was not fair use, distinguishing generative training [13]. Outside the United States, a Munich court largely ruled for the music rights society GEMA against OpenAI over memorised song lyrics, in a judgment that is not final [14]; the English High Court dismissed Getty Images' secondary infringement claim against Stability AI, holding the model not to be an infringing copy [15]. The US Justice Department filed a non-binding statement urging that training on written works is fair use [16].

Table 3. Legal records by channel and kind

Channel	Lawsuits	Rulings	Settlements	Regulatory actions	Guidance	All
1. Collection without a licence	35	25	7	1	2	70
6. Users' own data	3	2	0	7	0	12
Set aside (display, antitrust)	0	1	0	0	0	1
2 to 5 (paid channels)	0	0	0	0	0	0
All	38	28	7	8	2	83

Source: computed from [1] with the classification in Section 2.2. A case can contribute several records (a filing, a ruling, a settlement).

Within channel 1, 55 records are copyright claims; by content, news (23), music (11) and books (10) lead. OpenAI is a defendant in 20 records, Anthropic in 13 and Perplexity in 7. The seven settlement records cover four disputes: Bartz and three music cases.

3.2 Licensing from publishers and platforms

Volume and disclosure. The deal ledger records 111 disclosed licences to AI developers from July 2023 to 15 October 2026, of which 19 carry a dollar total or annual amount [1]. Table 4 shows who licenses and for what.

Table 4. Disclosed licensing deals by licensor type and use

Licensor type	Deals	Training	Both	Display or grounding	Unstated
News publisher	52	2	9	36	5
Music	18	13	0	0	5
Stock media	9	6	1	2	0
Academic publisher	7	5	0	0	2
Data marketplace	6	6	0	0	0
Social platform	4	0	2	1	1
Video library	4	4	0	0	0
Code or Q&A	3	1	1	1	0
Other, healthcare, archive, book	8	6	0	0	2
All	111	43	13	40	15

Source: computed from [1].

The table separates two products. News publishers mostly license content to be shown and cited in AI answers (36 of 52 deals), not to train on. Music, stock media, academic publishers and data marketplaces license mostly for training. Licensees are concentrated: OpenAI is party to 29 deals, Meta to 14, Google to 12, Microsoft to 7 and Amazon to 5.

Scale and structure. Reddit's prospectus disclosed \$203.0 million of aggregate contract value in data licensing agreements of two to three years [17]; its other revenue, mostly content licensing, was \$15.2 million in 2023, \$114.7 million in 2024 and \$140.0 million in 2025 [18]. News Corp's licence to OpenAI "could be" close to \$250 million over five years, as reported from The Wall Street Journal [19]. Several licensors report revenue that does not recur at the same level. Informa reported more than \$75 million of non-recurring data access revenue from AI partnerships in 2024 [20]. Wiley's AI licence revenue was \$13.7 million in the quarter to 31 July 2026, against \$28.9 million a year earlier [21]. Shutterstock's data, distribution and services revenue, which includes licences to AI developers, was \$203.3 million in 2025 [22] but \$56.1 million in the second quarter of 2026 against \$67.2 million a year earlier [23].

Litigation becomes licensing. Each of the three music settlements in the legal record is matched by a licence in the deal ledger dated the same day: Universal Music Group's settlement with Udio came with a licensed creation platform [24], and Warner Music Group settled with Udio [25] and with Suno, the latter announced as a partnership [26]. Here, channel 1 disputes resolved into channel 2 contracts.

Collective and display schemes. Google's News AI pilot pays more than 400 participating publications for "the delivery and use of enhanced content" [27]. Such schemes extend licensing below the largest publishers, mostly for display.

Exposure. For the developer, a licence replaces copyright exposure with contract terms. For the licensor, exposure shifts to whether it holds the rights it grants, which depends on its agreements with authors and users. Not legal advice.

3.3 Commissioned human and expert work

Scale. The best-documented money in the market flows to vendors that recruit people to make data to order. Table 5 lists the newest headline figure for the largest of them.

Table 5. Newest reported revenue or run rate of human-data vendors

Vendor	Figure	Measure and period	Basis
Mercor	over \$2 billion	gross revenue run rate, June 2026	reported [28]
Surge AI	\$1.2 billion	revenue, 2024	reported [29]
micro1	over \$1 billion	annual run rate, October 2026	company statement [30]
Handshake	over \$1 billion	gross annualised revenue, 2026	self-reported [31]
Scale AI	\$870 million	revenue, 2024	reported [32]
Snorkel AI	\$375 million	annualised run rate, September 2026	stated [33]
Innodata	\$251.7 million	revenue, 2025	SEC filing [34]
Appen	\$230.8 million	revenue, 2025	ASX filing [35]
TaskUs (AI Services segment)	\$214.2 million	revenue, 2025	SEC filing [36]

Source: compiled in [1]; each figure cited to its original. Measures differ and are not additive (Section 2.4).

The workforce and its pay. Handshake described paying about \$60 million a month to more than 30,000 experts [37]. Mercor says nearly five million experts are on its platform [38]. In the price compilation, 105 posted hourly rates for people who make training data have a median of \$77.46 (middle half \$54.55 to

\$101.34) [1]; Mercor's homepage stated an average contracted rate of \$112 an hour on 17 October 2026 [39]. The channel's lower end is documented too: TIME reported Kenyan workers on an OpenAI contract taking home \$1.32 to \$2 an hour in 2023 [40].

Concentration. Demand comes from few buyers. One customer accounted for about 58 percent of Innodata's 2025 revenue [41]. Chinese demand is reported as well: the six largest Chinese AI labs spend about \$500 million a year with American data-labelling companies, according to two founders citing estimates from Tencent and ByteDance executives, as reported by Forbes [42]. Capital has followed: Meta invested in Scale AI in June 2025, in a deal reported at about \$14.3 billion for a 49 percent stake [43]. Excluding that investment and acquisitions, funding rounds in the ledger with a stated amount total \$1,160 million across 12 rounds in 2024, \$1,042 million across 16 in 2025 and \$1,862 million across 30 in 2026 to 14 October [1]; recorded acquisitions rose from 4 in 2024 and 5 in 2025 to 13 in 2026.

Exposure. Legal exposure sits mainly with vendors, as employers or contractors of contributors and as holders of their personal data. Mercor's notice of a security incident is one example of the second [38]. Labour law is outside the scope of this paper.

3.4 Purchase of organisations' records

A new channel. Internal records show how work is done over years, which is what agentic models are trained to imitate. They sit behind logins that no crawler reaches. Public programmes to buy them appear in the record from 2026. Of the 37 buyers and intermediaries on the compiler's buyer board, 24 take company records [1]. micro1 committed \$1 billion over twelve months to buy and license enterprise operational data, financed by capital from Citi and Hercules Capital [44].

Transactions. The record holds one fully public sale. In the Chapter 11 case of Spirit Airlines, Google opened at \$5 million for the de-identified corporate archive and won at \$10 million; Mercor offered \$7 million and later \$10 million on condition of running de-identification with its own tools, and neither offer was selected; micro1 later filed a \$12.5 million proposal [45][46][47]. The asset schedule included about 100 million emails and 500 million Teams items [46]. Elsewhere, micro1 told the court it had completed more than 50 data transactions in 45 days, naming no sellers or prices [47]. A wind-down service reported nearly 100 deals recovering about \$1 million in its first year [48], then over 150 deals by October 2026, with its chief executive saying that 47 percent of its deals now include code or data sales [49]. Turing said it bought five to ten failed-startup codebases at an average in the tens of thousands of dollars [47]. In the price compilation, 93 published offers to holders have a median of \$316,228 per company, while four completed sales with a stated amount have a median of \$28,311 [1].

Supply. The pool of potential holders is large. In 2022 the United States had 561,160 employer firms with 20 to 99 employees, 93,341 with 100 to 499 and 21,041 with 500 or more [50]. In 2025, 81.5 percent of contributions on GitHub were made in private repositories [51].

Exposure. This channel's records are full of third parties: customers, employees and suppliers. In the Spirit sale, the consumer privacy ombudsman described data derived from about 190 million flights and touching the personal data of about 97 million passengers [52]; structured passenger systems were then removed from the sale. More than 100 members of Congress wrote to the chief executives of both companies asking them to halt the deal [53], and the approval hearing was moved to 6 November 2026 [54]. When Character.AI filed for Chapter 11 in October 2026, a state complaint in the case counted 21.6 million US

accounts, and what would happen to their conversations became part of the coverage [55]. The US Federal Trade Commission has stated that adopting more permissive data practices "through a surreptitious, retroactive amendment" to terms of service may be unfair or deceptive [56]. Customers are also writing training into contracts: language prohibiting AI training on customer data appeared in under 1 percent of cloud service agreements on one standard form in 2024, 11 percent in 2025 and 14 percent in 2026, across 16,140 agreements [57]. In the legal record, this channel's friction appears through bankruptcy procedure and political scrutiny rather than litigation. Not legal advice.

3.5 Built environments and synthetic data

What is bought. A reinforcement learning environment is simulated software with tasks and an automatic grader, in which a model practises by trial and error. Epoch AI's interviews found environment contracts often worth six to seven figures per quarter, about \$20,000 per website replica and about \$300,000 for a Slack-grade clone, and tasks at \$200 to \$2,000, with \$20,000 for one task rare but possible [58]. SemiAnalysis reported that OpenAI bought hundreds of website replicas [59]. Mercor's chief executive described frontier labs that "might buy 50,000 tasks a month" at \$2,000 per task to scale an order, against a cost basis of \$1,500 for a task of ten expert hours [60]. In the price compilation, buyers pay a median of \$2,837 per task made to order against \$300 paid to the people who make tasks, a ratio of 9.5 [1].

Supply and decay. Finished environments are scarce in the free supply: only 148 of the roughly 1.09 million datasets on the Hugging Face Hub carry the rl-environment tag [61]. Bounty programmes add some: Prime Intellect crowdsourced more than 400 environments in two months, paying \$100 to \$500 for open bounties and about \$1,000 to \$5,000 or more for application-only tasks [62]. Value decays quickly: one investor's analysis puts an environment's half-life at about six months and finds that the same customer rarely buys the same environment twice [63].

Synthetic data. Published research reports generation costs far below human collection. One study reports \$0.031 per synthetic demonstration against \$0.80 per human-annotated trajectory in an earlier benchmark [64], a ratio of about 26.

Exposure. Built data carries the least copyright exposure, because the developer or vendor makes it. Exposure moves to the inputs: several buyers turn organisations' anonymised records into simulated copies of those organisations, which ties channel 5 to channel 4 [1]. Not legal advice.

3.6 Users' own data

The regulatory path. Large platforms hold their users' posts, profiles and conversations already; using them for training requires a change of terms rather than a crawler or a licence. In the European Union this has run through the Irish Data Protection Commission as lead supervisor. Meta paused training on EU adults' public Facebook and Instagram posts in June 2024 [65] and resumed from 27 May 2025 after changes, under monitoring [66]. X agreed to suspend processing EU users' posts to train Grok in August 2024 [67]. LinkedIn began training on EU members' data from November 2025 after narrowing the data used and improving opt-out notices [68]. Twelve of the 83 legal records concern this channel, seven of them regulatory actions (Table 3). Private suits are newer: in October 2026, a Dutch consumers' association sued Meta over AI camera glasses [69], and proposed class actions in Illinois alleged that delivery robots captured biometric data later used for AI training [70].

A paid variant. Some companies now pay their own users for data. GoPro's programme had received more than 500,000 hours of subscriber video by its fourth-quarter 2025 report, with subscribers to earn half of the licence revenue [71]. A recording app paid US\$3 per approved hour and collected 38,000 hours from 3,900 contributors for a robotics company [72]. This variant sits between channels 3 and 6.

4. The shift, and what it changes for organisations that hold data

Table 6 summarises the six channels on the record.

Table 6. The six channels on the public record, October 2026

Channel	What developers obtain	Scale on the record	Price on the record	Who is paid	Main legal exposure (not legal advice)
1. Collection without a licence	Public pages, posts, books, code	2.17 billion pages in one monthly archive	No fee; one settlement of \$1.5 billion	Nobody, unless after a suit	Copyright, contract, anti-scraping; 70 of 83 legal records
2. Licensing	Archives and feeds, for training or display	111 disclosed deals, 19 priced	\$203.0 million aggregate (Reddit)	Publishers and platforms	Licensor's own rights
3. Commissioned work	Answers, rankings, tasks	Vendor run rates above \$1 billion	Median posted pay \$77.46 an hour	Contributors, through vendors	Employment and worker data
4. Organisations' records	Email, chat, tickets, code	24 of 37 buyers take them	Offers median \$316,228; one sale at \$10 million	The organisation holding them	Privacy, employees, third parties
5. Environments and synthetic data	Simulated software, generated data	Contracts of six to seven figures a quarter	\$2,837 median per task made to order	Builders	Mostly in the inputs
6. Users' own data	Posts, conversations, recordings	Platform-wide	No price, or a revenue share	Usually nobody	Data protection; 12 of 83 records

Sources: Sections 3.1 to 3.6.

Three readings follow. First, the shift is one of spending, not yet of volume. The open web remains the largest source by volume, and nearly nine in ten census sites still admit every named AI crawler. What has changed is where money goes: almost none to channel 1, and growing sums to channels 2 to 5. Second, the legal record and the money are almost disjoint. Disputes concentrate where nobody is paid (channels 1 and 6); the paid channels generate contracts, procedural scrutiny and policy letters rather than suits. Third, the paid channels reinforce one another. Vendors that commissioned experts to recreate how work is done have begun buying the records of how it was done, and turning those records into environments.

For organisations that hold data, four consequences follow from the record.

Access control moves value. Content a crawler can collect carries no licence fee unless the holder refuses access and negotiates a licence, and the disclosed licences in the record are with large publishers and platforms. Records behind a login are what channels 4 and 5 now pay for, and published offers for them run to six figures per company, although few completed sales are disclosed.

Rights decide what can be sold. The constraints visible in the Spirit sale (customer data, employee messages, third parties' confidential material) apply to any organisation's archive. So do the promises made when data was collected, the regulator's position on retroactive changes of terms, and the customer contracts that increasingly prohibit training on customer data.

Revenue may not recur. Several licensors report that AI licensing revenue fell after large initial deals, and environment buyers rarely buy the same environment twice. For holders, this suggests that the first sale to each buyer may carry much of the value.

There is more than one kind of buyer. A publisher's archive suits a licensing team; a software company's code and tickets may suit an environment builder; a services firm's correspondence may suit a records programme. The record shows these buyers pricing on different units and terms, which makes a single quoted figure a weak guide to value.

5. Limitations

The paper rests on a public-record compilation, with the biases that implies. Disclosure is selective: large licences surface through filings and the press, small sales rarely surface at all, and only 19 of 111 deals carry a dollar figure. Many figures are self-reported by vendors and intermediaries with an interest in appearing large, and several press figures rest on unnamed sources; both are marked where used, but cannot be checked here.

The census measures stated refusal, not crawling. A robots.txt rule does not show whether a crawler complied, and a site may block crawlers by other means. The current pass covers three of 46 slices, in the lower part of the ranked list; shares for the most visited sites may differ, and the earlier C4 audit suggests they are higher among the most actively maintained sources.

The channel boundaries overlap. Vendors that commission experts also build environments and buy records; consumer apps that pay users sit between channels 3 and 6; licences settle litigation. The classification of legal records is the author's, and other rules would move some records. Vendor figures mix gross and net revenue, run rates and annual results, and are not comparable across vendors.

The record is a snapshot read on 18 October 2026. Pending cases, an unapproved sale and programmes announced weeks earlier may change materially.

Finally, the author leads a company that acts as a seller's agent in this market and compiles the datasets used. That interest could bias which sources were sought and how they were classified. The datasets are public, each row links to its source, and the classification rule is stated so that others can reproduce or revise it.

6. Conclusion

The question of where AI training data comes from now has six answers rather than one. The open web remains the volume base but is increasingly refused, contested and exhausted at the margin; its legal cost has been priced by a billion-dollar settlement. Developers pay instead for licences, for people's expertise, for organisations' records and for built environments, and they draw on their own users under regulators'

supervision. The paid channels are where the growth in money is, and they are almost absent from litigation, which concentrates where nobody is paid.

For organisations that hold data, this shift turns an archive from a by-product into a potential asset whose value depends on access, rights, linkage and timing. It also places on them the obligations that come with any transfer of records about other people. The public record of these transactions is still thin, and its largest gap, the prices of completed sales, is the one that most limits what anyone can say about value.

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 compiles the datasets and census analysed here.

Data availability

The datasets are published as JSON under CC BY 4.0 at <https://aidata.broker/datasets/> (identifiers supply, deals, cases, financials, funding, market-log, buyers and prices), with browsable views at <https://aidata.broker/supply/>, [/deals/](https://aidata.broker/deals/), [/market-log/](https://aidata.broker/market-log/) and [/cases/](https://aidata.broker/cases/), and the census method at <https://aidata.broker/ai-crawlers/>. A narrative guide on the same records is at <https://aidata.broker/guides/how-ai-companies-get-training-data/>.

References

- [1] AIData.Broker, public datasets of the AI training data market, CC BY 4.0, each row with its source URL, dates and the source's words, reviewed 16 and 17 October 2026: supply (122 rows), deals (111), cases (83), financials (227), funding (111), market-log (110), buyers (37) and prices (1,022), at <https://aidata.broker/datasets/.json> (accessed 18 October 2026).
- [2] AIData.Broker, "Which sites block AI crawlers: a daily census of robots.txt," measured 17 October 2026. <https://aidata.broker/ai-crawlers> (accessed 18 October 2026).
- [3] Common Crawl, "September 2026 crawl archive now available." <https://commoncrawl.org/blog/september-2026-crawl-archive-now-available> (accessed 18 October 2026).
- [4] Epoch AI, "Will we run out of data? Limits of LLM scaling based on human-generated data," June 2024. <https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data> (accessed 18 October 2026).
- [5] Stack Exchange API 2.3, questions on stackoverflow created 1 to 30 September 2026 (total filter). <https://api.stackexchange.com/2.3/questions?fromdate=1788220800&todate=1790812800&site=stackoverflow&filter=total> (accessed 18 October 2026).
- [6] S. Longpre et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons," arXiv:2407.14933, 2024. <https://arxiv.org/abs/2407.14933> (accessed 18 October 2026).
- [7] Cloudflare, "The 2025 Cloudflare Radar Year in Review." <https://blog.cloudflare.com/radar-2025-year-in-review/> (accessed 18 October 2026).

- [8] B. Sookman, "Ziff Davis v. OpenAI: key copyright litigation ruling," 22 December 2025.
<https://barrysookman.com/2025/12/22/ziff-davis-v-openai-key-copyright-litigation-ruling/> (accessed 18 October 2026).
- [9] Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal.), Order on Fair Use, Dkt. 231, 23 June 2025.
<https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0.pdf> (accessed 18 October 2026).
- [10] Authors Guild, "Court grants final approval of Anthropic copyright settlement," July 2026.
<https://authorsguild.org/news/court-grants-final-approval-anthropic-copyright-settlement/> (accessed 18 October 2026).
- [11] Bartz v. Anthropic PBC, Plaintiffs' Updated Claims Report, Dkt. 643.
<https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.643.0.pdf> (accessed 18 October 2026).
- [12] Kadrey v. Meta Platforms, Inc. (N.D. Cal.), Order on Cross-Motions for Partial Summary Judgment, Dkt. 598, 25 June 2025.
<https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.598.0.pdf> (accessed 18 October 2026).
- [13] Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., No. 25-2153 (3d Cir.), opinion, 29 September 2026. <https://www2.ca3.uscourts.gov/opinarch/252153p.pdf> (accessed 18 October 2026).
- [14] Landgericht München I, press release on GEMA v. OpenAI, November 2025.
<https://www.justiz.bayern.de/gerichte-und-behoerden/landgericht/muenchen-1/presse/2025/11.php> (accessed 18 October 2026).
- [15] Getty Images v. Stability AI [2025] EWHC 2863 (Ch), 4 November 2025.
<https://caselaw.nationalarchives.gov.uk/ewhc/ch/2025/2863> (accessed 18 October 2026).
- [16] IPWatchdog, "DOJ sides with OpenAI, warns obstacles to AI development threaten national security," 3 September 2026. <https://ipwatchdog.com/2026/09/03/doj-sides-with-openai-warns-obstacles-to-ai-development-threaten-national-security/> (accessed 18 October 2026).
- [17] Reddit, Inc., Prospectus (SEC Form 424B4), March 2024.
<https://www.sec.gov/Archives/edgar/data/1713445/000162828024012380/reddit-final424b4.htm> (accessed 18 October 2026).
- [18] Reddit, Inc., Form 10-K for 2025 (SEC).
<https://www.sec.gov/Archives/edgar/data/1713445/000171344526000022/rddt-20251231.htm> (accessed 18 October 2026).
- [19] VentureBeat, "OpenAI partners with Wall Street Journal publisher News Corp," 22 May 2024.
<https://venturebeat.com/ai/openai-partners-with-wall-street-journal-publisher-news-corp> (accessed 18 October 2026).
- [20] Informa PLC, Annual Report and Accounts 2024.
<https://www.informa.com/globalassets/documents/investor-relations/2025/2024-informa-annual-report.pdf>

(accessed 18 October 2026).

[21] John Wiley & Sons, Form 10-Q for the quarter ended 31 July 2026 (SEC).

<https://www.sec.gov/Archives/edgar/data/107140/000162828026060632/jwa-20260731.htm> (accessed 18 October 2026).

[22] Shutterstock, Inc., Form 10-K for 2025 (SEC).

<https://www.sec.gov/Archives/edgar/data/1549346/000154934626000008/sstk-20251231.htm> (accessed 18 October 2026).

[23] Shutterstock, Inc., Form 10-Q for the quarter ended 30 June 2026 (SEC).

<https://www.sec.gov/Archives/edgar/data/1549346/000154934626000029/sstk-20260630.htm> (accessed 18 October 2026).

[24] Universal Music Group, "Universal Music Group and Udio announce Udio's first strategic agreements for new licensed AI music creation platform," 29 October 2025.

<https://www.universalmusic.com/universal-music-group-and-udio-announce-udios-first-strategic-agreements-for-new-licensed-ai-music-creation-platform/> (accessed 18 October 2026).

[25] WinBuzzer, "Warner Music Group and Udio settle copyright lawsuit," 20 November 2025.

<https://winbuzzer.com/2025/11/20/warner-music-group-and-udio-settle-copyright-lawsuit-as-suno-raises-250m-to-fight-on-xcxwn/> (accessed 18 October 2026).

[26] Warner Music Group, "Warner Music Group and Suno forge groundbreaking partnership," 25

November 2025. <https://www.wmg.com/news/warner-music-group-and-suno-forge-groundbreaking-partnership> (accessed 18 October 2026).

[27] Google, "News AI pilot: update on publishers," The Keyword, 15 October 2026.

<https://blog.google/products-and-platforms/products/news/news-ai-pilot-update-publishers/> (accessed 18 October 2026).

[28] The Information, "Mercor hit \$2 billion gross annualized revenue," 2026.

<https://www.theinformation.com/briefings/exclusive-mercor-hit-2-billion-gross-annualized-revenue> (accessed 18 October 2026).

[29] Business Insider, AI training companies and startups, September 2025.

<https://www.businessinsider.com/ai-training-companies-startups-2025-9> (accessed 18 October 2026).

[30] micro1, end of Q3 2026 update, 12 October 2026. <https://www.micro1.ai/blog/q3> (accessed 18 October 2026).

[31] TIME, "TIME100 AI 2026: Garrett Lord," August 2026. <https://time.com/collection/time100-ai/2026/garrett-lord/> (accessed 18 October 2026).

[32] TechCrunch, "Meta reportedly in talks to invest billions of dollars in Scale AI," 8 June 2025 (citing

Bloomberg). <https://techcrunch.com/2025/06/08/meta-reportedly-in-talks-to-invest-billions-of-dollars-in-scale-ai/> (accessed 18 October 2026).

- [33] TechCrunch, "Snorkel AI triples valuation to \$3.5B as demand for AI training data booms," 22 September 2026. <https://techcrunch.com/2026/09/22/snorkel-ai-triples-valuation-to-3-5b-as-demand-for-ai-training-data-booms/> (accessed 18 October 2026).
- [34] Innodata Inc., earnings release for 2025 (SEC Form 8-K, Exhibit 99.1). https://www.sec.gov/Archives/edgar/data/903651/000110465926020514/tm265812d1_ex99-1.htm (accessed 18 October 2026).
- [35] 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).
- [36] TaskUs, Inc., Form 10-K for 2025 (SEC). <https://www.sec.gov/Archives/edgar/data/1829864/000182986426000012/task-20251231.htm> (accessed 18 October 2026).
- [37] Handshake, Engagement Manager job posting, November 2025. <https://jobs.ashbyhq.com/handshake/f2ed7bba-e887-474d-9bf2-894a47f0082f> (accessed 18 October 2026).
- [38] Mercor, "Update on Mercor security incident." <https://mercor.com/blog/update-on-mercor-security-incident/> (accessed 18 October 2026).
- [39] Mercor, home page (average contracted rate as shown on 17 October 2026). <https://www.mercor.com/> (accessed 18 October 2026).
- [40] TIME, "OpenAI used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic," January 2023. <https://time.com/6247678/openai-chatgpt-kenya-workers/> (accessed 18 October 2026).
- [41] Innodata Inc., Form 10-K for 2025 (SEC). <https://www.sec.gov/Archives/edgar/data/903651/000110465926020655/inod-20251231x10k.htm> (accessed 18 October 2026).
- [42] Forbes Colombia, Spanish edition of the Forbes article by Anna Tong on Chinese AI labs and US data vendors. <https://forbes.co/ia/otro-problema-de-silicon-valley-en-china-esta-entrenando-su-ia> (accessed 18 October 2026).
- [43] TechCrunch, "Scale AI confirms significant investment from Meta, says CEO Alexandr Wang is leaving," 13 June 2025. <https://techcrunch.com/2025/06/13/scale-ai-confirms-significant-investment-from-meta-says-ceo-alexandr-wang-is-leaving/> (accessed 18 October 2026).
- [44] micro1, "micro1 commits \$1B to enterprise data acquisition," 9 October 2026. <https://www.micro1.ai/blog/micro1-commits-1b-to-enterprise-data-acquisition> (accessed 18 October 2026).
- [45] In re Spirit Aviation Holdings, Inc., No. 25-11897 (Bankr. S.D.N.Y.), Dkt. 1463, Notice of Auction Results with sale agreements. <https://storage.courtlistener.com/recap/gov.uscourts.nysb.331039/gov.uscourts.nysb.331039.1463.0.pdf> (accessed 18 October 2026).

[46] 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).

[47] 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).

[48] 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).

[49] 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).

[50] U.S. Census Bureau, 2022 Statistics of U.S. Businesses annual data tables. https://www2.census.gov/programs-surveys/susb/tables/2022/us_state_6digitnaics_2022.xlsx (accessed 18 October 2026).

[51] GitHub, "Octoverse: a new developer joins GitHub every second as AI leads TypeScript to #1," 2025. <https://github.blog/news-insights/octoverse/octoverse-a-new-developer-joins-github-every-second-as-ai-leads-typescript-to-1/> (accessed 18 October 2026).

[52] In re Spirit Aviation Holdings, Inc., Dkt. 1684, report of the consumer privacy ombudsman. <https://storage.courtlistener.com/recap/gov.uscourts.nysb.331039/gov.uscourts.nysb.331039.1684.0.pdf> (accessed 18 October 2026).

[53] The Record (Recorded Future News), "Lawmakers warn of Google-Spirit AI training deal," October 2026. <https://therecord.media/lawmakers-warn-of-google-spirit-ai-training-deal> (accessed 18 October 2026).

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

[55] Tech Times, "Character.AI bankruptcy puts intimate chats of 21M users on auction block," 15 October 2026. <https://www.techtimes.com/articles/329052/20261015/characterai-bankruptcy-puts-intimate-chats-21m-users-auction-block-40-ags-warn.htm> (accessed 18 October 2026).

[56] US Federal Trade Commission, Technology Blog, "AI (and other) companies: quietly changing your terms of service could be unfair or deceptive," 13 February 2024. <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> (accessed 18 October 2026).

[57] Common Paper, "2026 SaaS contract benchmark report." <https://commonpaper.com/resources/2026-saas-contract-benchmark-report/> (accessed 18 October 2026).

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

[59] SemiAnalysis, "RL environments and RL for science: data foundries and multi-agent architectures." <https://newsletter.semianalysis.com/p/rl-environments-and-rl-for-science> (accessed 18 October 2026).

[60] 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).

[61] Hugging Face, datasets filtered by the rl-environment tag. <https://huggingface.co/datasets?other=rl-environment> (accessed 18 October 2026).

[62] Prime Intellect, "Scaling the environments program," October 2025. <https://www.primeintellect.ai/blog/scaling-environments-program> (accessed 18 October 2026).

[63] Next Signal, "Data is a great place to start an AI company." <https://nextsignalprediction.substack.com/p/data-is-a-great-place-to-start-an> (accessed 18 October 2026).

[64] Synatra (arXiv:2409.15637, v2), cost comparison with Mind2Web. <https://arxiv.org/abs/2409.15637> (accessed 18 October 2026).

[65] Data Protection Commission (Ireland), "DPC's engagement with Meta on AI," June 2024. <https://www.dataprotection.ie/en/news-media/latest-news/dpcs-engagement-meta-ai> (accessed 18 October 2026).

[66] Data Protection Commission (Ireland), "DPC statement on Meta AI," 21 May 2025. <https://www.dataprotection.ie/en/news-media/latest-news/dpc-statement-meta-ai> (accessed 18 October 2026).

[67] Data Protection Commission (Ireland), "DPC welcomes X's agreement to suspend its processing of personal data for the purpose of training AI tool Grok," 8 August 2024. <https://www.dataprotection.ie/en/news-media/press-releases/dpc-welcomes-xs-agreement-suspend-its-processing-personal-data-purpose-training-ai-tool-grok> (accessed 18 October 2026).

[68] Data Protection Commission (Ireland), "DPC statement on LinkedIn AI training," November 2025. <https://www.dataprotection.ie/en/news-media/latest-news/dpc-statement-linkedin-ai-training> (accessed 18 October 2026).

[69] NL Times, "Dutch consumers' org also sues Meta over AI glasses; hearing set for November 26," 15 October 2026. <https://nltimes.nl/2026/10/15/dutch-consumers-org-also-sues-meta-ai-glasses-hearing-set-november-26> (accessed 18 October 2026).

[70] Chicago Sun-Times, "Lawsuits accuse Chicago delivery robot companies of collecting biometric data," 16 October 2026. <https://chicago.suntimes.com/news/2026/10/16/lawsuits-accuse-chicago-delivery-robot-companies-of-collecting-biometric-data> (accessed 18 October 2026).

[71] GoPro, Inc., Q4 and full-year 2025 management commentary. https://s21.q4cdn.com/291350743/files/doc_financials/2025/q4/GPRO-2025-12-31-Management-

Commentary.pdf (accessed 18 October 2026).

[72] The Globe and Mail, "Startups are hiring low-paid gig workers to train robots on manual work," 2026. <https://www.theglobeandmail.com/business/article-startups-are-hiring-low-paid-gig-workers-to-train-robots-on-manual/> (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). How AI developers acquire training data: six channels and the shift away from the open web. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-channels>