

# Who sells AI training data: a census of vendors, what they say they sell, and where their data comes from

Preprint

Yuma Heymans

AIData.Broker, <https://aidata.broker> | [hello@aidata.broker](mailto:hello@aidata.broker) | 18 October 2026

---

## Abstract

Firms that sell training data, expert labour, reinforcement-learning (RL) environments and evaluations to AI developers are mostly private and known largely through their own websites. This paper reports a census of 150 such vendors, read on their sites on 15 and 16 October 2026 with every claim tied to a quoted sentence, and links it to three companion tables: 37 organisations that publish a way to sell them data, 227 revenue records and 111 funding records. Evaluations are the most common offering (122 vendors, 81.3%), followed by datasets (104), RL environments (89), expert and human data (83) and annotation (57). Vendors bundle: the median vendor lists three of six kinds, and 79 of the 89 RL-environment sellers also sell evaluations. Most data are made to order: 90 vendors cite experts and contractors, 41 recordings, 30 generation and 15 software replicas, against 17 that license data from its owners and 22 that use the client's own data; 19 do not say. Of the 37 buyers, 24 take company records; among those that publish them, the median floor is \$100,000 and the median ceiling \$1 million per seller. Revenue disclosure runs from audited filings of four listed vendors to self-reported run rates, and 61 funding rounds with a disclosed amount total \$4.30 billion. We describe the method, the classification and the limits of a self-description census.

**Keywords:** AI training data; data vendors; data annotation; reinforcement learning environments; model evaluation; data markets; company data; market census

---

## 1. Introduction

Developers of AI models buy much of what they train and test on. They pay people to write, rank and check examples, buy datasets, license archives, and increasingly buy simulated software environments in which agents practise tasks, together with the evaluations that score them. A supply side has grown to meet that demand. Its firms range from listed companies with audited accounts to start-ups founded in 2026, and they describe themselves in overlapping terms: "data", "evals", "environments", "experts".

Two questions are hard to answer from the outside. What do these firms actually sell, once their self-descriptions are read side by side? And where does their data come from: is it made to order by contractors, generated, recorded, or licensed from people and companies that already hold it? The second question matters to anyone who holds data, to the people whose work or records end up in training sets, and to policymakers concerned with provenance.

This paper answers both descriptively, from a register of 150 vendors compiled from their own pages, and adds what three companion tables show about the market around them: the buyers that publish terms for acquiring data, and the revenue and funding the vendors have disclosed. All four tables are open under CC BY 4.0 [1, 2, 3, 4].

Five findings stand out.

1. Evaluations are the most common thing sold (122 of 150 vendors), and RL environments are sold by 89. These two travel together: 79 of the 89 RL-environment vendors also sell evaluations.
2. Vendors are multi-product. Only 15 list a single kind of offering; the median lists three.
3. Most vendors make data rather than acquire it. Of 131 that state a source, 87 name only made-to-order sources (experts, recordings, generation, replicas), and 44 name at least one existing source (licensed data, the client's data or public data).
4. A market for existing company records has appeared beside the vendors. Of 37 buyers with a public route in, 24 take company records, and 16 of the 150 vendors run such a programme themselves.
5. The money is visible unevenly. Four listed vendors file audited revenue; the largest private vendors are known through press reports and their own run-rate claims, which differ in period, basis and measure.

## 2. Data and method

### 2.1 The four tables

Table 1 lists the tables used. All are compiled by the author's organisation from public sources, published as JSON and CSV, and licensed under CC BY 4.0 for the compilation, with quotes keeping their sources' terms [1, 2, 3, 4].

**Table 1. Tables analysed**

| Table          | Rows | Reviewed    | Unit                                                   | Basis of rows                             |
|----------------|------|-------------|--------------------------------------------------------|-------------------------------------------|
| Vendors [1]    | 150  | 15 Oct 2026 | One vendor: what it sells, areas, sources, with quotes | The vendor's own pages                    |
| Buyers [2]     | 37   | 16 Oct 2026 | One organisation with a public way to sell it data     | 36 stated on its own page, 1 reported     |
| Financials [4] | 227  | 17 Oct 2026 | One revenue, run-rate or other figure for a period     | 195 stated, 16 reported, 16 self-reported |
| Funding [3]    | 111  | 17 Oct 2026 | One round, acquisition, investment or valuation        | 87 stated, 23 reported, 1 self-reported   |

Source: [1, 2, 3, 4].

**Vendors.** A row is a company that sells training data, RL environments, evaluations, expert work or annotation to AI developers [1]. Each row holds a one-line description in the compiler's words; a quoted sentence from the vendor's site that says what it does; a list of what it sells, each item with its own quote and page; its areas of work; a list of where its data comes from, each with a quote; a quote showing that it pays for outside data, where one exists; and, for 44 vendors, a headquarters and founding year with a quote. The register holds 475 quoted claims about what is sold and 223 about sources, on 319 distinct pages. Pages were read on 15 October 2026 (141 vendors) and 16 October 2026 (9). The register's page states the rule: each company "was read on its own site, and a script checked each quote word for word against the page; a claim whose quote did not match is left out" [5]. It also states that no vendor pays to be listed or chooses what its row says [5].

**Buyers.** A row is an organisation that publishes a way for others to sell or license it data: what it buys, what it says it pays (with a low and high figure where published), when it pays, eligibility rules, the kinds

of data it takes, any fee for introductions, and a note on what third parties have confirmed about it [2]. Pages were read between 11 and 16 October 2026.

**Financials and funding.** Financial rows record revenue, run rates, annual recurring revenue (ARR), guidance, customer concentration and similar figures for 30 companies; 174 of the 227 rows concern four listed companies (Appen, Innodata, TaskUs and TELUS Digital) [4]. Funding rows record 111 rounds, acquisitions, investments, financings, valuations and secondary sales from 19 January 2022 to 14 October 2026 [3]. Each row names its source, its date and its basis: stated by the company or a filing, reported by press, or self-reported (the company's own claim without other support).

Twenty-six of the 150 vendors have financial rows and 44 have funding rows (45 have either).

## 2.2 Classification

The register records what each vendor sells in six kinds and where its data comes from in seven. This paper uses those fields as recorded, and derives three further categories (Table 3 in Section 3.2):

- **Licensed corpora:** a vendor that sells datasets and says its data is licensed from its owners.
- **Company records:** a vendor that also appears on the buyer table with a programme that takes company records.
- **Pays for outside data:** a vendor that quotes a way to pay for outside data, or that appears on the buyer table (29 vendors; this matches the register page's own count [5]).

A **profile** assigns each vendor one primary model by a fixed rule: environments and evaluations only; expert data together with environments or evaluations; expert data without them; annotation and datasets without expert data; or datasets with environments or evaluations but no expert data. Data sources are grouped as **made to order** (made by experts and contractors, recorded from people and the real world, generated, replicas of real software) or **existing** (licensed from its owners, the client's own data, public data).

## 2.3 Computation and verification

All counts were computed from the JSON files downloaded on 18 October 2026, with a short Python 3 script. Shares are of 150 vendors unless stated. No figure is estimated or converted between currencies.

On 18 October 2026 the author re-fetched every quoted page in the vendor register and searched it for the quote. Of 865 quoted claims (150 descriptions, 475 offerings, 223 sources, 17 payments for data), 789 (91.2%) were found on a plain fetch of the cited page; 114 vendors had every quote found. Most of the remainder sit on pages that render their text with scripts. Three pages of one vendor now return "not found", two pages of another refuse automated requests with code 406, and one site refused with code 403. Of 72 rows checked among the financial rows of non-listed companies and the funding rounds of \$50 million or more, 57 quotes were found; most of the rest sit behind a paywall or a securities regulator's access rule, and two of those (Toloka's 2024 revenue and Mode's first-half 2026 revenue) were confirmed by fetching the filing with a declared user agent [21, 22].

## 3. Findings

### 3.1 What vendors sell

**Table 2. What vendors sell**

| Kind                       | Vendors | Share of 150 | Selling only this kind |
|----------------------------|---------|--------------|------------------------|
| Evaluations and benchmarks | 122     | 81.3%        | 1                      |
| Datasets                   | 104     | 69.3%        | 4                      |
| RL environments            | 89      | 59.3%        | 6                      |
| Expert and human data      | 83      | 55.3%        | 3                      |
| Annotation                 | 57      | 38.0%        | 1                      |
| Simulation                 | 20      | 13.3%        | 0                      |

Source: computed from [1]. A vendor can sell several kinds.

The 150 vendors list 475 offerings, a mean of 3.17 kinds each and a median of 3. Fifteen list one kind, 36 two, 32 three, 44 four, 22 five and one all six. The most common bundles are datasets, evaluations, expert data and RL environments (17 vendors); the same four plus annotation (17); evaluations and RL environments alone (16); and annotation, datasets, evaluations and expert data (16).

Evaluations are almost universal because they attach to everything else. They are sold by 79 of the 89 RL-environment vendors, 86 of the 104 dataset vendors and 72 of the 83 expert-data vendors. RL environments and annotation sit at opposite ends: only 22 vendors sell both, while 45 of the 57 annotation vendors also sell expert data. Simulation is the narrowest kind (20 vendors), and every simulation vendor also sells evaluations.

The derived profiles summarise the same pattern. Expert data together with environments or evaluations is the largest group (73 vendors, 48.7%), followed by environments and evaluations only (28, 18.7%), datasets with environments or evaluations but no expert data (27, 18.0%), annotation and datasets without expert data (12, 8.0%) and expert data without environments or evaluations (10, 6.7%).

The vendors' own words show what the categories hold. An environment can be a replica of real software: one vendor says each of its environments is "an independently built simulation of the service it represents, not the provider's live service" [6], and another advertises "interactive replicas of enterprise and consumer apps such as Jira, Salesforce, and Zendesk" [7]. Expert data is described by who writes it: "Every task is written by PhD researchers, practicing lawyers, and senior engineers, not crowdsourced" [8].

### 3.2 The classification requested

Table 3 applies the categories of Section 2.2.

**Table 3. Vendors by category**

| Category              | Rule applied to the register                                 | Vendors | Share |
|-----------------------|--------------------------------------------------------------|---------|-------|
| Evaluations           | Sells evaluations or benchmarks                              | 122     | 81.3% |
| RL environments       | Sells RL environments                                        | 89      | 59.3% |
| Expert work           | Sells expert and human data                                  | 83      | 55.3% |
| Annotation            | Sells annotation                                             | 57      | 38.0% |
| Simulation            | Sells simulation                                             | 20      | 13.3% |
| Licensed corpora      | Sells datasets and says its data is licensed from its owners | 15      | 10.0% |
| Company records       | Runs a published programme that takes company records        | 16      | 10.7% |
| Pays for outside data | Quotes a way to pay for data, or is on the buyer table       | 29      | 19.3% |

Source: computed from [1, 2].

The two smallest categories are the ones that depend on data someone else already holds. Fifteen vendors sell licensed corpora, among them creative and multimodal libraries licensed from contributors (Wirestock, Kled, Luel), medical data from data partners (Gradient Health), financial data (Dissei) and companies' internal records (Oak, Centific, Petrarch, The General Data Company) [1]. Sixteen run a programme to buy company records, and the overlap between those and the large expert-data vendors is direct: the buyer table's company-records programmes include those of micro1, Handshake AI, Mercor, Turing, Scale AI, Labelbox, SuperAnnotate, Appen and Centific [2]. One vendor's page puts the model plainly: "Oak Data licenses real business data, anonymizes it, and builds AI training environments" [9]. Another describes buying the data and working "with top AI labs to understand how real companies do work" [10].

### 3.3 Where the data comes from

**Table 4. Sources of data, as the vendors state them**

| Source                                  | Vendors | Share of 150 | Only this source |
|-----------------------------------------|---------|--------------|------------------|
| Made by experts and contractors         | 90      | 60.0%        | 33               |
| Recorded from people and the real world | 41      | 27.3%        | 9                |
| Generated (synthetic)                   | 30      | 20.0%        | 4                |
| The client's own data                   | 22      | 14.7%        | 6                |
| Licensed from its owners                | 17      | 11.3%        | 3                |
| Replicas of real software               | 15      | 10.0%        | 3                |
| Public data                             | 8       | 5.3%         | 2                |
| Not stated on its pages                 | 19      | 12.7%        |                  |

Source: computed from [1]. A vendor can state several sources.

Of the 131 vendors that state a source, 60 name one, 54 two, 13 three and 4 four. Eighty-seven name only made-to-order sources. Forty-four name at least one existing source, but only 12 name existing sources alone, and 25 name a source held by a third party (licensed or public data). The client's own data is a different case from licensing: the vendor works on data the buyer supplies, as when one vendor offers "a simulated copy of your systems, data, APIs, and users" [11].

Table 5 sets sources against what is sold.

**Table 5. Sources of data by what is sold (number of vendors)**

| What it sells         | Vendors | Experts | Recorded | Generated | Client's data | Licensed | Replicas | Public | Not stated |
|-----------------------|---------|---------|----------|-----------|---------------|----------|----------|--------|------------|
| Evaluations           | 122     | 81      | 30       | 28        | 19            | 13       | 14       | 7      | 13         |
| Datasets              | 104     | 73      | 36       | 22        | 15            | 15       | 9        | 6      | 6          |
| RL environments       | 89      | 50      | 16       | 24        | 17            | 10       | 14       | 4      | 13         |
| Expert and human data | 83      | 74      | 31       | 17        | 10            | 10       | 5        | 3      | 2          |
| Annotation            | 57      | 43      | 27       | 9         | 10            | 8        | 1        | 2      | 2          |
| Simulation            | 20      | 11      | 7        | 6         | 3             | 0        | 4        | 1      | 2          |

Source: computed from [1].

RL-environment vendors are the least likely to say where their data comes from (13 of 89 do not) and the most likely to generate it (24) or replicate software (14). Recorded data is concentrated among dataset and annotation vendors, and 26 of the 41 vendors that record data work in robotics, where data is captured with cameras and sensors: one vendor describes a capture app that "records demonstration sessions" [12]. Public data is rare, and sometimes stated in order to exclude other sources: one security firm says its environments "are derived exclusively from open-source software, and no customer data or security researchers are used at any stage of the training process" [13].

### 3.4 Areas, age and location

The register tags 142 vendors with at least one area. Coding is the most common (64 vendors, 42.7%), then enterprise work (59), finance (48), robotics (47), vision (45), medical (42), computer use (34), science (30), legal (29), speech (27), multilingual work (25), security (23), games (12), mathematics (12) and creative work (8).

A headquarters and a founding year are recorded for 44 vendors each, mostly from start-up directory listings (34 of the founding years). In that subset, 31 give San Francisco as their base and four more another Bay Area location, and 35 were founded in 2024 (9), 2025 (14) or 2026 (12). This subset is not representative of the register: it over-represents young, accelerator-backed firms, whose directory pages state these facts, and under-represents older firms whose sites do not.

### 3.5 The buyers' public programmes

Thirty-seven organisations publish a way to sell them data (Table 6). Eighteen buy for their own account, usually to turn what they buy into training data or environments for AI labs; twelve are marketplaces that pay a share of each sale; three are intermediaries, two are wind-down services for closing companies, and two pay for work to order.

**Table 6. Buyers by kind**

| Kind                               | Buyers | Publish a dollar figure | Take company records | Publish structured introducer terms |
|------------------------------------|--------|-------------------------|----------------------|-------------------------------------|
| Buys for its own account           | 18     | 11                      | 17                   | 9                                   |
| Marketplace, pays a share of sales | 12     | 1                       | 2                    | 0                                   |
| Intermediary                       | 3      | 1                       | 3                    | 1                                   |
| Wind-down service                  | 2      | 1                       | 2                    | 0                                   |
| Pays for work to order             | 2      | 2                       | 0                    | 0                                   |
| All                                | 37     | 16                      | 24                   | 10                                  |

Source: computed from [2].

**What they take.** Company records are the most common kind (24 buyers), followed by source code (13), video (10), images (9), audio and speech (8), datasets and environments (7), text archives (6), expert work (3) and physical-world data (3).

**Prices.** Sixteen buyers publish a dollar range or floor per seller (Table 7). Three more publish only a commission schedule, letting the seller set the price; four publish another figure (a rate per token, an average partner revenue, a range cited on their own pages, a reported range per book); and 14 publish none.

**Table 7. Buyers that publish a dollar figure, largest ceiling first**

| Buyer                  | Kind          | Low       | High                  | Source |
|------------------------|---------------|-----------|-----------------------|--------|
| Mode (BizBuyData)      | Own account   | \$100,000 | \$5,000,000           | [14]   |
| Handshake AI           | Own account   | \$100,000 | \$4,000,000           | [15]   |
| Polyshares             | Own account   | \$100,000 | \$2,000,000           | [16]   |
| micro1                 | Own account   | \$100,000 | \$1,000,000 and above | [17]   |
| Scale AI               | Own account   | \$10,000  | \$1,000,000 and above | [18]   |
| SuperAnnotate          | Own account   | \$100,000 | \$1,000,000 and above | [19]   |
| Turing                 | Own account   | \$10,000  | \$1,000,000           | [20]   |
| Ooak                   | Own account   | \$20,000  | \$1,000,000           | [9]    |
| Labelbox (Prism)       | Own account   | \$50,000  | \$1,000,000 and above | [23]   |
| HUMXN Research Labs    | Own account   | \$250,000 |                       | [24]   |
| Nyne                   | Own account   | \$100,000 |                       | [25]   |
| Troveo                 | Marketplace   | \$100,000 |                       | [26]   |
| Corpus Bridge          | Intermediary  | \$10,000  |                       | [27]   |
| SimpleClosure AssetHub | Wind-down     | \$10,000  |                       | [28]   |
| Prime Intellect        | Work to order | \$100     | \$5,000               | [29]   |
| Datacurve (Shipd)      | Work to order | \$80      | \$220                 | [30]   |

Source: [2] and the buyers' pages cited. Low and high are the buyer's published bounds; several are illustrative ("indicative", "up to", "could be worth").

Fourteen of these 16 take company records. Among company-records buyers, 14 publish a floor and 9 a ceiling; the median floor is \$100,000 and the median ceiling \$1,000,000. The two collectors pay for tasks rather than archives, in hundreds or low thousands of dollars per task [29, 30]. These are posted offers. In the buyer table's own notes, 16 of the 24 company-records programmes have no completed sale confirmed by a seller or by a third party [2]. One exception on the record is micro1, which said in a court filing that it had completed more than 50 data transactions in the preceding 45 days, "without identifying the sellers, prices, or AI-lab customers", as Fortune reported [31].

**Eligibility.** Rules are few. Six of the 24 company-records buyers publish a size rule: 20 or more employees (Handshake AI, Nyne, Oak, and Mode for office staff, with 10 or more for accounting firms and 6 or more for law firms), 30 or more (micro1) and 40 or more as a preferred start (Scale AI) [14, 15, 17, 18, 25, 9]. Two also require three years in operation [15, 25]. Nyne requires US employees [25], Mode wants teams primarily in the United States [14], and micro1 wants companies that work and document in English [17]. Eight take companies that are closing or have closed, among them Turing, Oak and the two wind-down services [2]. Twenty-five rows carry no structured criteria at all.

**Payment terms.** Table 8 groups what the buyers publish about when they pay.

**Table 8. When buyers say they pay**

| Payment terms                                          | Buyers |
|--------------------------------------------------------|--------|
| On delivery, acceptance or review of the data          | 9      |
| Per sale, or as the data licenses (a share of revenue) | 8      |
| On contract milestones                                 | 3      |
| Stated only as elapsed time                            | 3      |
| On completion of a task                                | 1      |
| Left to buyer and seller                               | 1      |
| Not published                                          | 12     |

Source: classified from [2].

Where a time is stated, it is short. Mercor pays by wire on delivery, which "typically occurs within 2-4 weeks of the signed agreement" [8]; Labelbox's page gives about 21 days to first wire [23]; Handshake AI's FAQ gives 7 to 14 business days for the whole process [2, 15]. Polyshares pays "Net 30 to 60 once the data is approved, anonymized, and delivered" [16], and Mode expects payment in about 60 days [2]. Several buyers tie payment to acceptance rather than delivery, as Appen does: "payment is on acceptance rather than submission" [32].

**Introductions.** Fifteen buyers publish something about paying for introductions, and ten in a structured form: a percentage with a cap (for example 5% of the deal up to \$100,000, in Mercor's referral terms as summarised in [2]), a flat sum (such as \$10,000 per successful referral [2, 15]), or the greater of a flat sum and a percentage [14]. All ten are buyers for their own account or intermediaries.

### 3.6 Revenue

Four vendors have filed audited accounts (Table 9). Three report AI-related revenue in the hundreds of millions of dollars; the fourth, TELUS Digital, is mainly a customer-experience business. The filings show dependence on a few customers.

**Table 9. Revenue of the listed vendors (US dollars, millions)**

| Company                     | FY2024 | FY2025                    | Concentration                                                           | Sources      |
|-----------------------------|--------|---------------------------|-------------------------------------------------------------------------|--------------|
| Appen                       | 234.3  | 230.8 (Appen China 102.9) | Google contract 82.8 in FY2023, 13.4 in FY2024                          | [33, 34, 35] |
| Innodata                    | 170.5  | 251.7                     | Largest customer about 48% (FY2024) and 58% (FY2025)                    | [36, 37, 38] |
| TaskUs, AI Services segment | 135.1  | 214.2                     | Segment of total service revenue of 995.0 (FY2024) and 1,183.5 (FY2025) | [39, 40]     |
| TELUS Digital, total        | 2,658  | taken private in 2025     | Google about 14.3% of FY2024 revenue                                    | [41, 42]     |

Source: [4] and the filings cited. TELUS Digital's revenue is mostly customer-experience services, not AI data.

The listed figures show growth at some vendors and fragility at others. Innodata's revenue rose from \$170.5 million to \$251.7 million between 2024 and 2025 while its largest customer's share rose from about 48% to about 58% [36, 37, 38]. Appen's revenue from its Google contract fell from \$82.8 million in 2023 to \$13.4 million in 2024 [33, 34]. TaskUs's AI Services segment grew from \$135.1 million to \$214.2 million [39, 40].

The larger private vendors are known through other channels (Table 10). Their figures differ in measure (revenue for a year, or a run rate that annualises a recent month), in period (from 2024 to October 2026), and in basis.

**Table 10. Newest headline revenue figure of private vendors**

| Company                | Figure                       | Measure and period                    | Basis         | Source |
|------------------------|------------------------------|---------------------------------------|---------------|--------|
| Mercor                 | over \$2 billion             | Gross run rate, June 2026             | reported      | [43]   |
| Surge AI               | \$1.2 billion                | Revenue, 2024                         | reported      | [44]   |
| micro1                 | over \$1 billion             | Run rate, October 2026                | stated        | [45]   |
| Handshake AI           | over \$1 billion             | Gross annualised revenue, August 2026 | self-reported | [46]   |
| Scale AI               | \$870 million                | Revenue, 2024                         | reported      | [47]   |
| Prolific               | set to surpass \$500 million | Annual revenue, May 2026              | reported      | [48]   |
| Snorkel AI             | \$375 million                | Run rate, September 2026              | stated        | [49]   |
| Turing                 | about \$300 million          | ARR, March 2025                       | stated        | [50]   |
| Proximal               | \$200 million                | Annualised revenue, September 2026    | self-reported | [51]   |
| Invisible Technologies | \$134 million                | Revenue, 2024                         | self-reported | [52]   |
| Prime Intellect        | over \$100 million           | Annualised revenue, July 2026         | self-reported | [53]   |
| Mecka AI               | over \$100 million           | Run rate, June 2026                   | self-reported | [54]   |
| AfterQuery             | over \$100 million           | Run rate, April 2026                  | self-reported | [55]   |
| Wirestock              | \$40 million                 | Run rate, May 2026                    | self-reported | [56]   |
| Toloka                 | \$26.4 million               | Revenue, 2024 (in a parent's filing)  | stated        | [21]   |

Source: [4] and the publications cited. Mercor's first-half 2026 gross revenue was reported at \$614 million [57].

Four private vendors report, or are reported at, \$1 billion or more. Seven of the 15 figures are the company's own claim without other support, four are press reports and four are stated by the company in an announcement or filing. Some are gross of what the vendor pays its contractors. This paper does not add them: a 2024 revenue figure and an October 2026 run rate are not one year of anything.

The growth series in the table are steep. micro1 reported \$7 million of ARR at the start of 2025 [58] and a run rate above \$1 billion in October 2026 [45]. Mercor reported \$75 million of ARR in February 2025 [59] and was reported at over \$2 billion of gross run rate in June 2026 [43].

### 3.7 Funding

Table 11 counts the funding records by kind of event.

**Table 11. Funding records by event**

| Event                                | Records | With a dollar amount | Sum of amounts   |
|--------------------------------------|---------|----------------------|------------------|
| Funding round                        | 69      | 61                   | \$4,297,419,397  |
| Acquisition                          | 23      | 4                    | \$2,469,000,000  |
| Investment                           | 7       | 7                    | \$14,361,803,000 |
| Financing (debt or equity programme) | 5       | 3                    | \$810,875,000    |
| Valuation (talks or announcement)    | 5       | 2                    | \$2,000,000,000  |
| Secondary sale                       | 2       | 2                    | \$175,000,000    |

Source: computed from [3]. Sums are arithmetic totals of stated amounts in US dollars and are not comparable across rows.

Funding rounds with a dollar amount total \$4,297,419,397 across 61 rounds; 59 completed rounds account for \$3,797,419,397 and two announced rounds for the rest. The median round is \$30 million. By year of the round, the totals are \$200 million in 2022 (one round), \$33.3 million in 2023, \$1.16 billion in 2024, \$1.04 billion in 2025 and \$1.86 billion in 2026 to 14 October (30 rounds with an amount). Twelve of the 69 rounds rest on press reports.

The largest rounds are Scale AI's \$1 billion Series F in May 2024 at a \$13.8 billion valuation [60], an announced \$400 million strategic seed for SignSplit at a stated \$1 billion valuation [61], Snorkel AI's \$350 million Series E at \$3.5 billion [62], Mercor's \$350 million Series C at \$10 billion [63] and Handshake's \$200 million Series F of January 2022 at \$3.5 billion, raised, in its words, to help Gen Z launch careers [64]. The investment total is dominated by one transaction: Meta's investment in Scale AI, reported at about \$14.3 billion [65]. Of the 23 acquisition records, three are the acquirer's own entry for a deal recorded again under the target, leaving 20 distinct acquisitions, of which four carry a price, the largest a reported talent-and-licence deal of about \$1.5 billion or more between Google and Mechanize [66].

## 4. Limitations

**Self-description.** The register records what vendors say, checked word for word against their pages; it does not test whether they deliver it. A vendor that lists evaluations may sell many or few. The categories are the vendors' language, mapped to six kinds by the compiler.

**Coverage.** The register is a census of the vendors the compiler found, not a sample from a defined population. Firms without English-language sites, firms that sell only through relationships, and in-house data teams at AI developers are under-represented. The register's page states no sampling frame.

**Absence of a claim.** Nineteen vendors do not say where their data comes from. "Not stated" is not evidence of any particular source, and a vendor that names one source may use others.

**Time.** Pages change. The register was read on 15 and 16 October 2026; by 18 October, three pages of one vendor no longer resolved. The buyers' terms also move: the buyer table records Oak raising its stated ceiling from \$300,000 to \$1 million [2]. Unlike the deals, buyers, funding and financials tables, the vendor register is not in the Zenodo snapshot of 13 October 2026 [67], so it has no DOI yet.

**Money.** Revenue figures mix audited revenue, gross run rates and self-reported claims, and the funding table mixes equity, debt and acquisition prices. Sums in Table 11 are arithmetic, not market sizes.

**Position of the compiler.** The author's organisation acts as an agent for organisations that sell data to buyers such as those in Section 3.5, and some intermediaries in the buyer table do similar work. That interest could affect what is collected and how it is classified. The quote on every claim and the open tables are meant to make that checkable.

Nothing in this paper is legal advice, and no row is a finding about any company's conduct.

## 5. Conclusion

Read side by side, 150 vendors' own descriptions show a market organised around bundles. Evaluations attach to almost everything, RL environments have become a mainstream product sold by 89 vendors, and most vendors sell three or more kinds of offering. Most data is made to order, by experts, recordings, generation or software replicas. A smaller group builds from data that already exists, and around it a set of 37 buyers now publishes routes for holders to sell, with company records the most common target and posted floors and ceilings that cluster around \$100,000 and \$1 million.

What the census cannot show is volume. The vendors' revenues are visible for four listed companies and, for the rest, through press reports and their own run-rate claims; the buyers' offers are visible, their completed purchases mostly not. The next useful step for the register is fields that separate a vendor's capability claims from evidence that a buyer has paid for them.

## **Data availability**

The tables are available under CC BY 4.0 at <https://aidata.broker/datasets/vendors.json>, <https://aidata.broker/datasets/buyers.json>, <https://aidata.broker/datasets/funding.json> and <https://aidata.broker/datasets/financials.json> (CSV alongside). The versions this paper uses, downloaded on 18 October 2026, are deposited together as a snapshot of that day at <https://doi.org/10.5281/zenodo.23452930>, the first snapshot to include the vendor register. The analysis script is available from the author on request.

## **Conflict of interest**

The author is chief executive of AIData.Broker, which compiles and publishes the tables analysed here and acts as an agent for organisations that sell or license data to AI developers. Some organisations in the buyer table are potential counterparties of the author's organisation, and some intermediaries in it offer similar services.

## **Funding**

The tables and this analysis were produced by AIData.Broker. No other funding was received.

## **Use of AI tools**

Generative AI tools were used to assist with the scripted analysis and the drafting of this manuscript. The author reviewed the text, the tables and the sources and takes responsibility for the content.

## **References**

All URLs were accessed on 18 October 2026 unless a note says otherwise.

[1] AIData.Broker, "AI training data vendors" (dataset, JSON; reviewed 15 October 2026), CC BY 4.0. <https://aidata.broker/datasets/vendors.json>.

[2] AIData.Broker, "Buyers of AI training data" (dataset, JSON; reviewed 16 October 2026), CC BY 4.0. <https://aidata.broker/datasets/buyers.json>.

- [3] AIData.Broker, "Funding of AI data companies" (dataset, JSON; reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/funding.json>.
- [4] AIData.Broker, "Revenue of AI data companies" (dataset, JSON; reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/financials.json>.
- [5] AIData.Broker, "Who sells AI training data, and where each gets it" (web page, updated 15 October 2026). <https://aidata.broker/vendors>. Related pages: "Who buys AI training data, and what each one publishes," <https://aidata.broker/buyers>; "The AI training data market, in numbers," <https://aidata.broker/market>; guides "AI training data companies: who does what in 2026," <https://aidata.broker/guides/ai-training-data-companies>, and "micro1, Handshake, Turing, Mercor: data programs compared," <https://aidata.broker/guides/company-data-buyer-programs-compared>.
- [6] Archal, "Stateful API sandboxes for AI agent testing" (home page). <https://archal.ai/>.
- [7] Turing, "RL environments: UI, MCP and backend environments for agents." <https://www.turing.com/advance/rl-environments>.
- [8] Mercor, "Off-the-shelf AI training datasets (licence ready)." <https://www.mercor.com/apex/off-the-shelf-data/>. Data page: <https://mercor.com/data>.
- [9] Oak, partner page. <https://www.oakdata.com/partner>.
- [10] Handshake AI, data partnerships page. <https://joinhandshake.com/ai/data-partnerships>.
- [11] Veris AI, home page. <https://veris.ai>.
- [12] 6thSense, privacy page. <https://6thsense.dev/privacy>.
- [13] Bugcrowd, "Bugcrowd launches reinforcement learning environments to help AI models learn real-world security skills" (press release), 21 May 2026. <https://www.bugcrowd.com/press-release/bugcrowd-launches-reinforcement-learning-environments-to-help-ai-models-learn-real-world-security-skills>.
- [14] BizBuyData (Mode), referral and seller page. <https://bizbuydata.com/50k>.
- [15] Handshake AI, data licensing page and FAQ. <https://joinhandshake.com/ai/data-partnerships/license>.
- [16] Polyshares, site and broker portal. <https://polyshares.com/>.
- [17] micro1, data partnerships page. <https://www.micro1.ai/data-partnerships>.
- [18] Scale AI, data partnership page. <https://scale.com/data-partnership>.
- [19] SuperAnnotate, data partnerships page. <https://www.superannotate.com/data-partnerships>.
- [20] Turing, Project Lazarus page. <https://lazarus.turing.com/>.
- [21] Nebius Group N.V., Annual report (SEC Form 20-F) for 2024, filed 30 April 2025. <https://www.sec.gov/Archives/edgar/data/1513845/000155837025005991/nbis-20241231x20f.htm>.

- [22] Mode Mobile, Inc., Semiannual report (SEC Form 1-SA), 9 October 2026.  
<https://www.sec.gov/Archives/edgar/data/1748441/000149315226046559/form1-sa.htm>.
- [23] Labelbox, Prism home page. <https://prism.labelbox.com/>.
- [24] HUMXN Research Labs, sell company data page. <https://humxn.io/sell-company-data>.
- [25] Nyne, sell company data page. <https://nyne.ai/sell-company-data/>.
- [26] Troveo, business data page. <https://www.troveo.ai/business-data>.
- [27] Corpus Bridge, home page. <https://corpusbridge.com/>.
- [28] SimpleClosure, "What are your company's assets worth?", 3 September 2026.  
<https://simpleclosure.com/blog/insights/what-your-assets-are-worth/>.
- [29] Prime Intellect, "Scaling our open-source environments program," 27 October 2025.  
<https://www.primeintellect.ai/blog/scaling-environments-program>.
- [30] Shipd (Datacurve), home page, read 16 October 2026. <https://shipd.ai/>.
- [31] Fortune, "Little-known AI startup Micro1 tries to trump Google's bid for bankrupt Spirit Airlines' data," 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/>.
- [32] Appen, enterprise data page. <https://www.appen.com/enterprise-data>.
- [33] Appen Limited, ASX announcement, full-year 2023 results, 27 February 2024.  
<https://www.asx.com.au/asxpdf/20240227/pdf/060v9j6h0ljb6l.pdf>.
- [34] Appen Limited, ASX announcements, full-year 2024 results, 26 February 2025.  
<https://announcements.asx.com.au/asxpdf/20250226/pdf/06fyjdg1v66466.pdf>;  
<https://www.asx.com.au/asxpdf/20250226/pdf/06fyj3zjx089dd.pdf>.
- [35] Appen Limited, ASX announcement, full-year 2025 results, 25 February 2026.  
<https://announcements.asx.com.au/asxpdf/20260225/pdf/06wq7ntr51bhgl.pdf>.
- [36] Innodata Inc., Fourth quarter and full-year 2024 results (SEC Form 8-K, Exhibit 99.1), 20 February 2025. [https://www.sec.gov/Archives/edgar/data/903651/000110465925015700/tm255624d1\\_ex99-1.htm](https://www.sec.gov/Archives/edgar/data/903651/000110465925015700/tm255624d1_ex99-1.htm).
- [37] Innodata Inc., Fourth quarter and full-year 2025 results (SEC Form 8-K, Exhibit 99.1), 26 February 2026. [https://www.sec.gov/Archives/edgar/data/903651/000110465926020514/tm265812d1\\_ex99-1.htm](https://www.sec.gov/Archives/edgar/data/903651/000110465926020514/tm265812d1_ex99-1.htm).
- [38] Innodata Inc., Annual reports (SEC Form 10-K) for 2024, filed 24 February 2025, and for 2025, filed 26 February 2026. <https://www.sec.gov/Archives/edgar/data/903651/000141057825000194/inod-20241231x10k.htm>; <https://www.sec.gov/Archives/edgar/data/903651/000110465926020655/inod-20251231x10k.htm>.
- [39] TaskUs, Inc., Annual report (SEC Form 10-K) for 2024, 6 March 2025.  
<https://www.sec.gov/Archives/edgar/data/1829864/000182986425000016/task-20241231.htm>.

- [40] TaskUs, Inc., Annual report (SEC Form 10-K) for 2025, 5 March 2026.  
<https://www.sec.gov/Archives/edgar/data/1829864/000182986426000012/task-20251231.htm>.
- [41] TELUS International (Cda) Inc. (TELUS Digital), Fourth quarter 2024 earnings release (SEC Form 6-K), 13 February 2025.  
[https://www.sec.gov/Archives/edgar/data/1825155/000162828025005301/q4\\_2024xearningsxrelease.htm](https://www.sec.gov/Archives/edgar/data/1825155/000162828025005301/q4_2024xearningsxrelease.htm).
- [42] TELUS International (Cda) Inc. (TELUS Digital), Annual report for 2024, 13 February 2025.  
<https://www.sec.gov/Archives/edgar/data/1825155/000162828025005299/tixt-20241231.htm>. Take-private agreement: TELUS Corporation and TELUS Digital (SEC Form 6-K, Exhibit 99.1), 2 September 2025.  
[https://www.sec.gov/Archives/edgar/data/1825155/000110465925086091/tm2524985d1\\_ex99-1.htm](https://www.sec.gov/Archives/edgar/data/1825155/000110465925086091/tm2524985d1_ex99-1.htm).
- [43] The Information, briefing on Mercor's gross annualised revenue, 6 July 2026 (title not verified).  
<https://www.theinformation.com/briefings/exclusive-mercor-hit-2-billion-gross-annualized-revenue> (paywalled; not opened by the author; figure as recorded in [4]).
- [44] Business Insider, "Meet the startups turning human data into billion-dollar businesses," 25 September 2025. <https://www.businessinsider.com/ai-training-companies-startups-2025-9>.
- [45] micro1, "micro1 end of Q3 update," 12 October 2026. <https://www.micro1.ai/blog/q3>.
- [46] TIME, "TIME100 AI 2026: Garrett Lord," 27 August 2026. <https://time.com/collection/time100-ai/2026/garrett-lord/>.
- [47] TechCrunch, "Meta reportedly in talks to invest billions of dollars in Scale AI," 8 June 2025.  
<https://techcrunch.com/2025/06/08/meta-reportedly-in-talks-to-invest-billions-of-dollars-in-scale-ai/>.
- [48] Business Post, "Irish-founded AI firm Prolific set to surpass \$500m in annual revenue," 15 May 2026 (body behind a paywall; headline read). <https://www.businesspost.ie/tech/irish-founded-ai-firm-prolific-set-to-surpass-500m-in-annual-revenue/>.
- [49] 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/>.
- [50] TechCrunch, "Turing, a key coding provider for OpenAI and other LLM producers, raises \$111M at a \$2.2B valuation," 6 March 2025. <https://techcrunch.com/2025/03/06/turing-a-key-coding-provider-for-openai-and-other-llm-producers-raises-111m-at-a-2-2b-valuation/>.
- [51] Proximal, "New frontiers," 29 September 2026. <https://www.proximal.ai/blog/new-frontiers/>.
- [52] Invisible Technologies, "\$100M fundraise to power the next gen of AI infrastructure," 16 September 2025. <https://invisibletech.ai/blog/invisible-technologies-raises-100-million>.
- [53] Prime Intellect, "\$130M Series A to build the open superintelligence stack," 8 July 2026.  
<https://www.primeintellect.ai/blog/series-a>.
- [54] Mecka AI, "Mecka raises \$60M Series B led by Sequoia Capital," 7 October 2026.  
<https://mecka.ai/news/series-b>.

- [55] AfterQuery, "Human expertise, reimagined," 9 April 2026. <https://www.afterquery.com/blog/human-expertise-reimagined>.
- [56] TechCrunch, "Wirestock raises \$23M to supply creative multimodal data to AI labs," 14 May 2026. <https://techcrunch.com/2026/05/14/wirestock-raises-23m-to-supply-multi-modal-data-to-ai-labs/>.
- [57] The Information, report on Mercor's revenue and customers, 21 July 2026 (title not verified). <https://www.theinformation.com/articles/mercors-fast-growth-relies-biggest-ai-companies-documents-show> (paywalled; the dataset notes that only the headline could be read).
- [58] TechCrunch, "micro1, a competitor to Scale AI, raises funds at \$500M valuation," 12 September 2025. <https://techcrunch.com/2025/09/12/micro1-a-competitor-to-scale-ai-raises-funds-at-500m-valuation/>.
- [59] TechCrunch, "Mercor, an AI recruiting startup founded by 21-year-olds, raises \$100M at \$2B valuation," 20 February 2025. <https://techcrunch.com/2025/02/20/mercor-an-ai-recruiting-startup-founded-by-21-year-olds-raises-100m-at-2b-valuation/>.
- [60] Scale AI, "Scale's Series F: expanding the data foundry for AI," 21 May 2024. <https://scale.com/blog/scale-ai-series-f>.
- [61] SignSplit (PR Newswire), "SignSplit secures \$400 million strategic seed round at \$1 billion valuation," 5 October 2026. <https://www.prnewswire.com/news-releases/signsplit-secures-400-million-strategic-seed-round-at-1-billion-valuation-302897779.html>.
- [62] Snorkel AI, "Snorkel AI raises \$350M to scale the data factory for frontier AI," 22 September 2026. <https://snorkel.ai/press/snorkel-ai-raises-350m-to-scale-the-data-factory-for-frontier-ai/>.
- [63] TechCrunch, "Mercor quintuples valuation to \$10B with \$350M Series C," 27 October 2025. <https://techcrunch.com/2025/10/27/mercor-quintuples-valuation-to-10b-with-350m-series-c/>.
- [64] Handshake (PR Newswire), "No network required: Handshake raises \$200M to help Gen Z launch careers," 19 January 2022. <https://www.prnewswire.com/news-releases/no-network-required-handshake-raises-200m-to-help-gen-z-launch-careers-301463445.html>.
- [65] 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/>.
- [66] AI Weekly (relaying Business Insider), "Google closes \$1.5B Mechanize talent deal, Epoch AI cofounder joins DeepMind midtraining," 11 September 2026. <https://aiweekly.co/alerts/google-closes-15b-mechanize-talent-deal-epoch-ai-cofounder-joins-deepmind>.
- [67] AIData.Broker, "The AI training data market: prices, deals, buyers and cases (snapshot of 18 October 2026)" [Data set], Zenodo, 2026. <https://doi.org/10.5281/zenodo.23452930>.

## 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.

Licence: CC BY 4.0. Cite as: Heymans, Y. (2026). Who sells AI training data: a census of vendors, what they say they sell, and where their data comes from. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-vendors>