

AI data licensing deals, July 2023 to October 2026: a public dataset and what it shows

Data Descriptor with descriptive analysis

Yuma Heymans

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

Abstract

Licences of content to developers of artificial intelligence (AI) are negotiated privately and recorded in no registry; what is known comes from announcements, securities filings and press reports. This paper describes an open dataset that compiles those disclosures, with a source link, date and verbatim quote for every row, and reports what it shows. The version analysed here (reviewed 17 October 2026) holds 111 records dated 11 July 2023 to 15 October 2026. It names 24 licensees; OpenAI appears in 29 records, Meta in 14 and Google in 12. News publishers are the licensor in 52 records (46.8%) and music companies in 18. The use licensed shifted: 66.7% of 2024 records include training, against 36.1% in 2025 and 44.7% in 2026 to date, while licences to show content in AI answers became the most common news arrangement. Disclosure of money is sparse and skewed. Nineteen records (17.1%) carry a dollar figure, but only 11 of the 89 records that describe a specific licence do, and 7 of those 11 rest on press reports citing unnamed people. No music licence in the table has a stated amount. A contract term is stated in 7 records. We describe the fields, collection and dating rules, versioning and verification, and the disclosure and survivorship biases that limit inference. The dataset is released under CC BY 4.0, with a snapshot on Zenodo.

Keywords: AI training data; content licensing; data licensing; generative AI; disclosure bias; news publishers; music licensing; open dataset

1. Introduction

Since mid-2023, publishers, platforms, music companies and image libraries have licensed content to companies that build AI systems. The licences cover two different products. Some permit training: the licensee may use the content to fit the parameters of a model. Others permit display or grounding: the licensee may retrieve the content at the moment a user asks a question and show, quote or summarise it in the answer, usually with a link. Many licences cover both, and some announcements do not say which.

There is no public register of these agreements. Terms are confidential by default, the parties choose what to announce, and the record is scattered across newsrooms, securities filings, earnings calls and trade press. Researchers, policymakers and rights holders each rebuild the same list, often from secondary summaries that drop the source and blur what a party stated with what a reporter heard.

This paper describes a dataset built to remove that duplication: one row per disclosed arrangement, with its parties, content, use, term and value where stated, the disclosing publication, its date and the source's own words. The dataset is maintained by the author's organisation and is published under a Creative Commons Attribution 4.0 licence [1, 5]. The same organisation has published narrative guides on parts of this record, on Reddit's licences, on news publishers and on the size of the market [6, 7, 8]; this paper is the first full description of the table itself.

The paper documents the table's fields, rules, versions and verification, then reports who licenses what, for which use, over time, and how much money is visible.

Five findings stand out.

1. Licensees are concentrated: 24 are named, OpenAI is party to 29 of 111 records, and OpenAI, Meta, Google and Microsoft together appear in 58.
2. News is the largest licensor group, with 52 records. Music follows with 18, all dated between August 2025 and October 2026.
3. The share of records that include training fell after 2024, from 66.7% of 2024 records to 36.1% in 2025. In news, 18 of 21 records dated 2025 license display or grounding only.
4. Money is rarely disclosed. Nineteen records carry a dollar figure, but those 19 figures are of at least six different kinds (a licence total, an annual fee, a licensor's annual revenue from all licensees, a programme budget, a pool, a court-auction bid) and cannot be summed. Among records of a specific licence, 11 of 89 carry a figure, and 7 of those 11 are press reports from unnamed sources.
5. The record moves: between the 13 October snapshot and the 17 October review, 16 rows were added (12 dated before October 2026) and 6 corrected.

2. Data and method

2.1 Scope and unit of record

A record is a licence, sale or paid access arrangement for data or content, made to an organisation that develops or deploys AI, and disclosed in a public source on or after July 2023 [1].

The usual unit is one announced arrangement between one licensor and one licensee. Several records depart from that unit, and the departures are part of the data:

- Five records are aggregates, flagged in the content field with the word "AGGREGATE": a licensor's total from all its licensees, such as Reddit's aggregate contract value for the licences it signed in January 2024 [9] or Shutterstock's 2023 AI licensing revenue [11].
- Eleven records describe a programme, pool or marketplace with many licensors, such as Google's News AI pilot [42, 49], Microsoft's Publisher Content Marketplace [48], Perplexity's revenue-share programmes [18, 19] and micro1's budget to buy company data [29].
- Four records are listings or rate cards offered to any buyer, one records a negotiation without a confirmed deal [30], and one records a winning bid at a court-supervised sale that the court had not approved when the table was reviewed [26].

The remaining 89 records describe a specific licence, or a set of licences announced together, by one licensor to one or more specific licensees, named or not. Table 7 in Section 3.5 gives the breakdown.

2.2 Fields

The dataset has 19 fields (Table 1); value fields are filled only where a source states a figure.

Table 1. Fields of the deals table

Field	Content
id	Stable slug
date	Date made or announced (YYYY-MM-DD, or YYYY-MM)
licensor, licensorId	Licensor as named; identifier in the companies table
licensorType	One of 12 categories (Table 5)
licensee, licenseeIds	Licensee as named; identifiers (empty when unnamed)
content	What was licensed (compiler's words; flags such as AGGREGATE)
use	training, display/grounding, both, or unstated
termYears	Contract length in years, where stated as a number
valueUsd, annualUsd	A total, or an amount per year, in US dollars, where stated
valueText	The value as described, or that it was not disclosed
unitPriceText	A price per unit (per image, per hour, per book), where stated
basis	stated (by a party, filing or named official) or reported (press citing unnamed people)
source, sourceLabel, sourceDate	URL, name and date of the publication relied on
quote	The supporting sentence from the source

2.3 Collection and dating rules

Records come from a party's own announcement, a regulatory filing (for example a prospectus or 8-K exhibit filed with the US Securities and Exchange Commission), an earnings-call transcript or named press. Aggregator lists help find candidates, but a row rests on the publication that carries the fact.

All 111 rows have a verbatim quote and a source URL, drawn from 81 distinct URLs and 59 publications. TechCrunch is the most frequent (26 rows), then BNN Bloomberg (7 rows from one article on Meta's news licences [41]). Twenty-seven rows cite a party's own announcement, a filing, an earnings-call transcript or a marketplace listing; 84 cite press.

The date is that of the arrangement or its announcement, not of reading; 105 rows have a full date and 6 only a month. Revenue aggregates are dated at the end of the period they report. For 83 of the 105 fully dated rows, the source is dated the same day as the record; for 10 it is dated more than 30 days later, the longest gap being 489 days, where the row for a licence of May 2024 relies on an article of September 2025 [15].

Of 111 rows, 93 are stated and 18 reported; all 18 reported rows cite press. Basis attaches to the row: a row can be stated (the parties announced the deal) while its value text attributes a figure to people familiar with it.

2.4 Identifiers and companion tables

Identifiers point to a companion table of 452 organisations with a kind and a description [2]. All 24 licensee identifiers and all 93 licensor identifiers (63 distinct licensors) resolve in it.

A second table, the brokerage log [3], holds 110 dated records of buyer programmes, price ranges, referral terms and transactions in sales of company records (2 October 2025 to 16 October 2026). Sales of

company records (email, chat, tickets, code) are mostly kept there, because they are rarely announced as deals between named parties.

2.5 Licence, versions and access

The tables are published as JSON and CSV at stable URLs [1, 2, 3] and on a web page [4]. The licence is CC BY 4.0 and, in the dataset's own words, it covers "our compilation of the tables"; "quotes keep their sources' terms, and article text is not covered" [1].

The version this paper analyses is deposited on Zenodo as a snapshot of 18 October 2026, DOI 10.5281/zenodo.23452930 [5], with 111 deal rows among 15 tables; an earlier snapshot of 13 October (DOI 10.5281/zenodo.23346240) held 95 deal rows. Compared by id, no row was removed and 16 were added; 12 of the additions are dated before October 2026 (the earliest April 2026), so they were found late rather than signed late. Six rows changed in substance: three dates, two licensor names, three value descriptions and one source (replaced with the buyer's own announcement). Citations of a figure should name the version used.

2.6 Derived variables and computation

Four variables used in Section 3 are derived by the author and are not fields of the dataset:

- **Content family** groups the 12 licensor types into six families (Table 5).
- **Record type** assigns each row to one of the six types in Section 2.1, using the AGGREGATE flag and the content description.
- **Kind of figure** classifies each of the 19 dollar figures by what it measures, from the source's words (Table 6).
- **Source type** separates a party's own announcement, filings, transcripts and listings from press.

All counts were computed from the JSON files with a short Python 3 script. Shares are rounded to one decimal. No figure is estimated, imputed or converted between currencies.

2.7 Verification for this paper

On 18 October 2026 the author fetched all 111 source URLs and searched each page for the row's quote. It was found for 103 rows (including four that cite an archived Reuters article [30]). Eight could not be checked because the publisher refused automated requests: six cite one Press Gazette article [27], one a blog [18] and one a company newsroom [20]. For two, the figure was confirmed elsewhere: Perplexity's pool in Axios [19] and the Tempus terms in Recursion's 8-K filing [21].

3. Findings

3.1 Over time

The 111 records run from Shutterstock's expanded licence to OpenAI on 11 July 2023 [37], followed two days later by the Associated Press's licence to OpenAI [40], to three records dated 15 October 2026, among them Google's update to its News AI pilot, which says more than 400 publications take part [42]. Table 2 counts records by year and use, and Table 3 by quarter.

Table 2. Records by year and use

Year	Records	Training only	Training and display	Display only	Use not stated	Share including training
2023 (from July)	4	3	1	0	0	100.0%
2024	33	13	9	8	3	66.7%
2025	36	12	1	19	4	36.1%
2026 (to 15 October)	38	15	2	13	8	44.7%
All	111	43	13	40	15	50.5%

Source: computed from [1]. "Display" covers display and grounding in AI answers.

Table 3. Records by quarter

Quarter	Records	With a dollar figure
2023 Q3	2	0
2023 Q4	2	1
2024 Q1	5	2
2024 Q2	16	4
2024 Q3	7	1
2024 Q4	5	1
2025 Q1	7	0
2025 Q2	2	0
2025 Q3	6	1
2025 Q4	21	0
2026 Q1	6	2
2026 Q2	10	2
2026 Q3	15	4
2026 Q4 (to 15 October)	7	1

Source: computed from [1]. Month-only dates are placed in the quarter of their month.

Two quarters stand out. The second quarter of 2024 holds 16 records, including Reddit's and News Corp's licences to OpenAI and four records from one Reuters article on image and video licensing [30]. The fourth quarter of 2025 holds 21: Meta's seven news licences of 5 December 2025 [41] and eight music records, among them the three major labels' licences to Klay Vision [47] and Universal Music Group's settlement and licences with Udio [46]. Music records fall in five of the six quarters from mid-2025 (2, 8, 0, 2, 5 and 1).

Quarterly counts measure announcements, not signings, and move with late finds: nine of the 16 rows added after the snapshot fall in the third quarter of 2026.

The shift in use is the clearest change. All four 2023 records include training and 22 of 33 in 2024 do; in 2025 the share falls to 36.1%, mostly because of news: of 21 news records dated 2025, 18 license display or grounding only, one training only and one both. OpenAI's own records follow the same path. Of its 13 records dated 2024, 8 cover training and display; of its 13 records dated 2025 and 2026, 7 cover display only, 2 training only and 4 do not state the use. The partial rebound in 2026 comes from outside news. Of the 17 records dated 2026 that include training, 5 have a music licensor, 5 a data marketplace, 4 concern company data, code or map data (such as micro1's purchasing programme [29] and Turing's purchases of failed start-ups' codebases [36]), 2 an image or video library and 1 a news publisher.

3.2 Licensees

The table names 24 licensees by identifier (Table 4). A record that names several licensees counts once for each: four records do so, among them Shutterstock's licences to Meta, Google, Amazon and Apple [30]. Eighteen records name no licensee or only "any buyer".

Table 4. Licensees named in three or more records

Licensee	Records	Training only	Training and display	Display only	Not stated	With a dollar figure
OpenAI	29	4	9	11	5	5
Meta	14	2	1	11	0	1
Google	12	4	1	6	1	2
Microsoft	7	3	0	4	0	0
Amazon	5	2	1	2	0	0
Udio	4	3	0	0	1	0
Perplexity	4	0	0	4	0	1
ElevenLabs	3	2	0	0	1	0
Suno	3	3	0	0	0	0
Stability AI	3	1	0	0	2	0
Klay Vision	3	3	0	0	0	0
Unnamed or any buyer	18	14	1	1	2	7

Source: computed from [1]. Protege and Apple appear in two records each, and eleven further licensees in one.

OpenAI's 29 records include 21 with news publishers. Meta's 14 include 12 with news publishers, of which 11 license display only; the seven announced on 5 December 2025 are described in the dataset as real-time news and links in the Meta AI chatbot [1, 41]. Google's 12 records spread across news (6), code, image and video licences, and a winning bid for a bankrupt airline's internal business data, pending court approval [26]. The music-model developers (Udio, Suno, Stability AI, Klay Vision, ElevenLabs) account for 16 records between them, all with music licensors, and at least one followed litigation: Universal Music Group's agreement with Udio pairs "the compensatory legal settlement" with new licences [46].

Records with unnamed licensees are the most informative about money (7 of 18 carry a dollar figure), because a licensor reporting to investors can disclose an amount without naming the buyer, as Reddit [9]

and John Wiley & Sons [23, 24] did.

3.3 Licensors and content

Table 5 counts records by licensor type, with the content family used in this paper.

Table 5. Records by licensor type and use

Licensor type	Family	Records	Includes training	Display only	Not stated	With a dollar figure
News publisher	News text	52	11	36	5	6
Music	Music and audio	18	13	0	5	0
Stock media	Images and video	9	7	2	0	2
Academic publisher	Books and journals	7	5	0	2	4
Data marketplace	Mixed and other	6	6	0	0	1
Other	Mixed and other	5	4	0	1	2
Social platform	User posts and code	4	2	1	1	3
Video library	Images and video	4	4	0	0	0
Code and Q&A	User posts and code	3	2	1	0	0
Healthcare data	Mixed and other	1	0	0	1	1
Government or archive	Mixed and other	1	1	0	0	0
Book publisher	Books and journals	1	1	0	0	0

Source: computed from [1]. "Includes training" counts training only plus training and display.

By family: news text 52, music and audio 18, images and video 13, mixed and other 13, books and journals 8, user posts and code 7.

Three patterns follow. First, display-only licensing is almost entirely a news phenomenon: 36 of the 40 display-only records have a news licensor. Second, every other family is mainly licensed for training: music, images and video, books and journals, and the mixed group together have 41 records that include training and 2 that are display only. Third, the families differ in what they disclose. Academic publishers, social platforms and image libraries disclose amounts more often than news (4 of 7, 3 of 4 and 2 of 9, against 6 of 52), largely through licensors' reports to investors. Music discloses no money: none of its 18 records has a dollar figure, a unit price or a licence amount in its text.

The content field shows the range: a news agency's archive back to 1983, queried by a chatbot [50], more than 200 Disney, Marvel, Pixar and Star Wars characters for video generation [38], and a bankrupt airline's internal email and operations data, passenger records excluded [1, 26].

3.4 How much of the money is visible

Nineteen records (17.1%) carry a dollar figure in the numeric fields: nine a total and ten an amount per year. Fourteen records carry a unit price, two of them also among the 19. Adding records whose value text states an amount without a numeric field, 36 records carry some monetary amount for the arrangement, and 75 carry none. The text-only amounts include ranges and approximations that the compiler did not reduce to a number: The New York Times's licence to Amazon at "\$20 million to \$25 million annually" according to The Wall Street Journal [43], Axel Springer's licence to OpenAI at "tens of millions of euros" per year according to the Financial Times [44], and Shutterstock's initial licences to large technology companies at "\$25 million to \$50 million each", in its chief financial officer's words to Reuters [30].

The 19 figures do not measure the same thing (Table 6). Only 11 describe a specific licence or a set of licences announced together. The others are a licensor's revenue from all its AI licensees (four records), a total contract value across unnamed licensees (one), a programme budget (one), a pool for a publisher programme (one) and a winning bid at a bankruptcy auction awaiting court approval (one).

Table 6. The 19 records with a dollar figure, by size

Date	Licensor to licensee	Figure	What the figure measures	Basis	Source
2026-10	Companies to micro1	\$1 billion	Buyer's programme budget for 12 months, not one deal	stated	[29]
2024-05	News Corp to OpenAI	about \$250 million	One licence, five-year total ("could be")	reported	[10]
2024-01	Reddit to unnamed licensees	\$203.0 million	Aggregate contract value, terms of two to three years	stated	[9]
2023	Shutterstock to several AI companies	\$104 million a year	Licensor's 2023 AI licensing revenue	stated	[11]
2024	Informa to several AI companies	over \$75 million a year	Licensor's 2024 revenue from AI partnerships	stated	[12]
2026-03	Getty Images to two unnamed licensees	about \$65 million	Combined value of two multiyear licences	stated	[13]
2024-05	Reddit to OpenAI	about \$60 million a year	One licence, annual fee	reported	[15]
2024-02	Reddit to Google	about \$60 million a year	One licence, annual fee	reported	[14]
2026-03	News Corp to Meta	up to \$50 million a year	One licence, annual ceiling, three years	reported	[16]
FY2026	John Wiley & Sons to several licensees	\$49 million a year	Licensor's fiscal 2026 AI revenue	stated	[17]
2025-08	Publishers to Perplexity	\$42.5 million	Pool for the first phase of a programme	stated	[18, 19]
2026-09	Tempus AI to Recursion	\$42 million	Committed payments over three years	stated	[20, 21]
2026-05	Wirestock to six model makers	\$40 million a year	Licensor's annual run-rate revenue	stated	[22]
2024-06	John Wiley & Sons to an unnamed tech company	\$23 million	One content rights project	stated	[23]
2024-09	John Wiley & Sons to a second unnamed tech company	\$21 million	One content rights project	stated	[24]
2024-05	Dotdash Meredith to OpenAI	at least about \$16 million a year	Fixed part of one licence	reported	[25]
2026-08	Spirit Airlines estate to Google	\$10 million	Winning auction bid, not yet approved	stated	[26]
2026-09	Bennett, Coleman and Co to OpenAI	around \$5 million a year	One licence, annual estimate	reported	[27, 28]
2026-09	Indian Express Group to OpenAI	about \$3 million a year	One licence, annual estimate	reported	[27, 28]

Source: [1] and the original publications cited. "Stated" and "reported" are the dataset's basis field.

Four observations follow from the table.

The visible money is mostly reported, not stated, at the level of a single licence. Of the 11 single-licence figures, 4 come from a party (Getty Images on an earnings call [13], Tempus AI [20], and Wiley's two content rights projects in filings [23, 24]) and 7 from press reports. The best known of them, about \$60 million a year from each of Google and OpenAI to Reddit, are press figures; Reddit's own filing states only the \$203.0 million aggregate and terms of "two to three years", with "a minimum of \$66.4 million of revenue to be recognized" in 2024 [9]. News Corp's licence to OpenAI is known through The Wall Street Journal's figure as relayed by VentureBeat, which says it "could be" close to \$250 million over five years [10]. The two Indian figures come from an Indian trade publication's estimate relayed by Press Gazette; the parties did not state them [27, 28].

The figures cannot be added. A five-year total, an annual ceiling, a licensor's yearly revenue from all buyers and a programme pool are different quantities, and they overlap (Reddit's January 2024 aggregate covers licences that include the one reported for Google [9, 14]). This paper reports no total.

Disclosure follows reporting to investors. Eight of the 19 figures come from companies speaking to investors or filing with a securities regulator: Reddit's prospectus [9], three figures from John Wiley & Sons' filings [17, 23, 24], Informa's annual report [12], Getty Images' earnings call [13], Shutterstock's chief executive on its AI licensing revenue [11], and the Tempus licence, whose terms the licensee also filed [21]. Companies that report to public investors describe material revenue; licensors without that obligation often do not. This may be one reason the music records carry no amounts.

Display licences are the least priced. Of the 40 display-only records, one has a numeric dollar figure (Perplexity's pool [18, 19]) and one more an amount in words (seventeen publishers' "six-figure deals" for enterprise retrieval through Snowflake, as reported by Digiday and relayed by Press Gazette [27]).

The unit prices are a separate and smaller window. Freepik licensed most of its 200 million images "at 2 to 4 cents per image" [30]; Defined.ai's chief executive cited "\$1 to \$2 per image, \$2 to \$4 per short-form video and \$100 to \$300 per hour of longer films", and "\$0.001 per word" for text [30]; the terms of Microsoft's three-year licence of HarperCollins titles reportedly offered "\$5,000 per nonfiction book, split evenly between the author and HarperCollins" [32]; Adobe paid contributors "about \$2.62 per minute of submitted video" on average [31]; unpublished creator video sold for \$1 to \$4 per minute [33]; and conversational speech lists at \$60 to \$95 per hour in one catalogue [34]. Company data sits at a different scale: defunct start-ups' archives sold through a wind-down service for \$10,000 to \$100,000 per company, as reported [35], and Turing's chief executive said it had paid an average in the tens of thousands of dollars for each failed start-up's codebase, as Fortune reported citing The Information [36].

3.5 What kind of record carries the money

Table 7 sets the record types of Section 2.1 against disclosure.

Table 7. Record types and disclosure of a dollar figure

Record type	Records	With a dollar figure	Stated	Reported
Specific licence (or set of licences announced together)	89	11	4	7
Programme, pool or marketplace with many licensors	11	2	2	0
Aggregate across licensees	5	5	5	0
Listing or rate card offered to any buyer	4	0	0	0
Winning bid at a court-supervised sale, not yet approved	1	1	1	0
Negotiation reported, no deal confirmed	1	0	0	0

Source: computed from [1]; record type derived as described in Section 2.6.

The aggregates are fully priced because they exist to report money. The specific licences, which are what most readers picture as "deals", are priced in 12.4% of cases (11 of 89), and only 4.5% (4 of 89) by a party. Within the 89 licences, 34 license display only, 29 training only, 13 both and 13 do not state the use.

3.6 Terms

A contract length is stated as a number in 7 records (Table 8).

Table 8. Records with a stated term

Date	Licensor to licensee	Term (years)	Use	Basis	Source
2023-07	Shutterstock to OpenAI	6	training	stated	[37]
2024-05	News Corp to OpenAI	5	training and display	reported	[10]
2024-11	HarperCollins to Microsoft	3	training	reported	[32]
2025-12	The Walt Disney Company to OpenAI	3	not stated	stated	[38]
2026-03	News Corp to Meta	3 (at least)	training and display	reported	[16]
2026-09	Tempus AI to Recursion	3 (payment period)	not stated	stated	[20, 21]
2026-10	Reach to Google	2	display	stated	[27]

Source: computed from [1].

For Tempus AI and Recursion, the three years are the period of the amended payments; Recursion's filing shows the amendment extended the agreement "from five to six years from its effective date" (November 2023), with fees of \$14,000,000 on each of the third, fourth and fifth anniversaries [21]. Reddit's licences run "two to three years" [9] and Informa's licence to Microsoft pays recurring fees "over the next three years" [45], but neither has a number in the term field. Thirteen more records say "multi-year" without a length, among them Agence France-Presse's licence to Mistral AI [50].

Seven terms are too few for a distribution. They are short (two to six years), and terms travel with figures: three of the seven term records carry a dollar figure, against 16 of the other 104.

3.7 Arrangements that are not cash for content

Some records show money moving the other way, or in kind. Disney's licence of characters to OpenAI carries no disclosed fee, and Disney makes "a \$1 billion equity investment in OpenAI" [38]. The major

labels licensed their catalogues to Stability AI "as part of the deal" in which they joined a funding round reported at \$76 million [39]. OpenAI funds Axios's expansion into four new markets [51], and Perplexity shares revenue rather than paying a fee [18, 19]. A count of deals is not a count of payments to licensors.

4. Limitations

Disclosure bias. The table holds what someone chose, or was obliged, to publish. Licensors with investors to inform, and licensees that want to be seen paying for content, announce; small licensors and buyers of company records mostly do not. News publishers, listed companies and the largest AI developers are therefore over-represented, and private archives and company data under-represented, as the dataset's web page itself notes [4].

Value bias. Filings disclose what is material to a listed licensor; press reports from unnamed people may describe a first year, a ceiling or a bundle of cash and services, and cannot be checked; a buyer's programme announces an offer, not a transaction. The 19 figures are not a sample of licence values in any statistical sense.

Survivorship and lag. Arrangements enter when found: 12 of the 16 rows added between 13 and 17 October 2026 are dated earlier, some by six months, so recent quarters are likely to grow. Arrangements that ended quietly or were not announced are absent. One gap: Recursion's 8-K for the 2026 amendment describes a master agreement of November 2023 with annual licence fees of \$22 million, \$32 million and \$42 million [21]; the deals table records the 2026 amendment but not the 2023 agreement as a separate row.

Unit of record. One article can produce several rows (Meta's seven licences of 5 December 2025 [41]), and one row can stand for many licences (Reddit's aggregate; Google's pilot with more than 400 publications [42]). Rows count disclosed arrangements as the compiler split them, not contracts.

Classification. "Unstated" use (15 records) means the source does not say, not that no training occurs. Licensor types, families and record types are judgments, stated in Section 2.6 so that others can change them.

Small numbers. With 111 rows and cells of 1 to 52, differences are descriptive; no significance test would be meaningful for a non-random record.

Position of the compiler. The compiler acts as an agent for organisations that license data to AI developers, an interest that could affect what is searched for and how rows are classified. Sources and quotes on every row, and published versions, are meant to make any such effect checkable.

Nothing in this paper is legal advice, and no row is a statement about the lawfulness of any party's conduct.

5. Conclusion

Compiled with a source and quote for every row, the public record of AI data licensing from July 2023 to October 2026 has a clear shape. A few AI developers, led by OpenAI, Meta and Google, are the licensees. News publishers are the largest group of licensors, and their licences moved from training towards display

in answers from 2025. Music became the second-largest group in late 2025, licensing for training and disclosing no money.

Little of the money is public. Seventeen per cent of records carry a dollar figure, those figures measure at least six different things, and for single licences most are press reports. Anyone citing a price for "an AI licensing deal" should say which kind of figure it is and whether a party stated it.

Three improvements to the dataset would help users most: a field for the kind of figure (Table 6 shows the categories), a field that marks programmes and aggregates without relying on the content text, and frozen deposits on Zenodo at each review so that every count in a paper can be tied to a DOI.

Data availability

The deals table, the companies table and the brokerage log are available as JSON and CSV at <https://aidata.broker/datasets/deals.json>, <https://aidata.broker/datasets/companies.json> and <https://aidata.broker/datasets/market-log.json> under CC BY 4.0. The snapshot of 13 October 2026 is archived at <https://doi.org/10.5281/zenodo.23346240>. The analysis in this paper uses the version reviewed on 17 October 2026, downloaded on 18 October 2026. 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 dataset described here and acts as an agent for organisations that license or sell data to AI developers. The organisations named in this paper are named because public sources reported their arrangements.

Funding

The dataset 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 data licensing deals" (dataset, JSON; reviewed 17 October 2026), CC BY 4.0. <https://aidata.broker/datasets/deals.json>.

[2] AIData.Broker, "Organisations in the AI data market" (dataset, JSON; reviewed 17 October 2026). <https://aidata.broker/datasets/companies.json>.

[3] AIData.Broker, "AI data brokerage log" (dataset, JSON; reviewed 17 October 2026). <https://aidata.broker/datasets/market-log.json>.

- [4] AIData.Broker, "AI data licensing deals: every disclosed deal since 2023" (web page, updated 17 October 2026). <https://aidata.broker/deals>.
- [5] 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>.
- [6] AIData.Broker, "Reddit's AI data deals: what Google and OpenAI pay" (guide, 17 October 2026). <https://aidata.broker/guides/reddit-ai-data-licensing-deals>.
- [7] AIData.Broker, "AI licensing deals with news publishers: every disclosed price" (guide, 17 October 2026). <https://aidata.broker/guides/news-publisher-ai-licensing-deals>.
- [8] AIData.Broker, "How big is the AI training data market? What the records show" (guide, 17 October 2026). <https://aidata.broker/guides/ai-training-data-market-size>.
- [9] Reddit, Inc., Prospectus (SEC Form 424B4), 21 March 2024. <https://www.sec.gov/Archives/edgar/data/1713445/000162828024012380/reddit-final424b4.htm>.
- [10] 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>.
- [11] Bloomberg Law, "Shutterstock's AI-licensing business generated \$104 million," 4 June 2024. <https://news.bloomberglaw.com/ip-law/shutterstocks-ai-licensing-business-generated-104-million>.
- [12] Informa PLC, Annual Report 2025, Chief Executive's review, 14 April 2026. <https://www.informa.com/globalassets/documents/investor-relations/2026/informa-annual-report-2025-ceo-review.pdf>.
- [13] The Motley Fool, "Getty Images (GETY) Q4 2025 earnings call transcript," 16 March 2026. <https://www.fool.com/earnings/call-transcripts/2026/03/16/getty-images-gety-q4-2025-earnings-transcript/>.
- [14] The Verge, "Google cut a deal with Reddit for AI training data," 22 February 2024. <https://www.theverge.com/2024/2/22/24080165/google-reddit-ai-training-data>.
- [15] Sherwood News, "Reddit bounces on report that it's in talks with Google, OpenAI on fresh data-sharing deal," 17 September 2025. <https://www.sherwood.news/business/reddit-bounces-on-report-that-its-in-talks-with-google-openai-on-fresh-data>.
- [16] Editor & Publisher, "News Corp, Meta in AI content licensing deal worth up to \$50 million a year," 4 March 2026. <https://www.editorandpublisher.com/stories/news-corp-meta-in-ai-content-licensing-deal-worth-up-to-50-million-a-year,260471>.
- [17] John Wiley & Sons, Inc., Fourth quarter fiscal 2026 results (SEC Form 8-K, Exhibit 99.1), 16 June 2026. <https://www.sec.gov/Archives/edgar/data/107140/000162828026043447/ex991fy26q4.htm>.
- [18] S. Palmer, "Perplexity will pay publishers," 26 August 2025. <https://shellypalmer.com/2025/08/perplexity-will-pay-publishers/> (refused automated access on 18 October 2026; figure confirmed in [19]).

- [19] Axios, "Perplexity says it will give publishers 80% cut of new subscription product's revenue," 26 August 2025. <https://www.axios.com/2025/08/26/perplexity-comet-plus-subscription>.
- [20] Tempus AI, "Tempus and Recursion extend existing data license agreement and enter new license agreement for Recursion's RNA foundation model," 21 September 2026. <https://www.tempus.com/news/pr/tempus-and-recursion-extend-existing-data-license-agreement-and-enter-new-license-agreement-for-recursions-rna-foundation-model/> (refused automated access on 18 October 2026; terms confirmed in [21]).
- [21] Recursion Pharmaceuticals, Inc., Current report (SEC Form 8-K), event of 15 September 2026. <https://www.sec.gov/Archives/edgar/data/1601830/000160183026000112/rxr-20260915.htm>.
- [22] 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/>.
- [23] John Wiley & Sons, Inc., Fourth quarter fiscal 2024 results (SEC Form 8-K, Exhibit 99.1), 13 June 2024. <https://www.sec.gov/Archives/edgar/data/107140/000010714024000104/wly-2024430xex991.htm>.
- [24] John Wiley & Sons, Inc., First quarter fiscal 2025 earnings presentation (SEC Form 8-K exhibit), 5 September 2024. <https://www.sec.gov/Archives/edgar/data/107140/000010714024000186/q125earningspresentation.htm>.
- [25] Adweek, "OpenAI is paying Dotdash Meredith at least \$16 million to license its content," 18 November 2024. <https://www.adweek.com/media/openai-dotdash-meredith-licensing-payment/>.
- [26] The Star (Reuters), "Google to buy Spirit Airlines business data for \$10 million," 17 August 2026. <https://www.thestar.com.my/tech/tech-news/2026/08/18/google-to-buy-spirit-airlines-business-data-for-10-million>. Status: TravelPulse, "Spirit Airlines' \$10 million data sale to Google delayed as lawmakers raise concerns," 10 October 2026. <https://www.travelpulse.com/news/airlines-airports/spirit-airlines-google-10-million-data-sale-delayed-as-lawmakers-raise-concerns>.
- [27] Press Gazette, article tracking news publishers' AI deals and lawsuits, as dated 12 October 2026 in the dataset (title not verified). <https://pressgazette.co.uk/platforms/news-publisher-ai-deals-lawsuits-openai-google/> (refused automated access on 18 October 2026; rows rest on the dataset's reading).
- [28] exchange4media, "OpenAI's TOI & IE deals spark wider interest among news publishers," 2026 (title as listed by a search engine). <https://www.exchange4media.com/digital-news/openais-toi-ie-deals-spark-wider-interest-among-news-publishers-158488.html> (refused automated access on 18 October 2026).
- [29] micro1, "micro1 announces commitment to spend \$1B on enterprise data acquisition over the next 12 months," 9 October 2026. <https://www.micro1.ai/blog/micro1-commits-1b-to-enterprise-data-acquisition>.
- [30] Reuters, "Inside Big Tech's underground race to buy AI training data," 5 April 2024 (archived copy). <https://gwern.net/doc/www/www.reuters.com/923c2769eb8628509a3827dc9204892fd2f22993.html>.
- [31] Maginative, "Adobe is paying \$3 per minute for videos to train AI model," 11 April 2024. <https://www.maginative.com/article/adobe-is-paying-3-per-minute-for-videos-to-train-ai-model>.

[32] eWeek, "HarperCollins book catalog to train Microsoft AI models for the next 3 years," 7 December 2024. <https://www.eweek.com/news/harpercollins-books-train-microsoft-ai-models/>.

[33] The Decoder, "OpenAI and Google are buying YouTubers' unpublished videos for up to \$4 per minute," 11 January 2025. <https://the-decoder.com/openai-and-google-are-buying-youtubers-unpublished-videos-for-up-to-4-per-minute/>.

[34] SpeechData.ai, "How to buy speech data: the complete buyer's guide," 18 June 2026. <https://www.speechdata.ai/guides/how-to-buy-speech-data>.

[35] Gizmodo, "Failed companies are selling old Slack chats and email archives to train AI," 17 April 2026. <https://gizmodo.com/failed-companies-are-selling-old-slack-chats-and-email-archives-to-train-ai-2000747916>.

[36] 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/>.

[37] TechCrunch, "Shutterstock expands deal with OpenAI to build generative AI tools," 11 July 2023. <https://techcrunch.com/2023/07/11/shutterstock-expands-deal-with-openai-to-build-generative-ai-tools/>.

[38] TechCrunch, "Disney signs deal with OpenAI to allow Sora to generate AI videos featuring its characters," 11 December 2025. <https://techcrunch.com/2025/12/11/disney-signs-deal-with-openai-to-allow-sora-to-generate-ai-videos-featuring-its-characters/>.

[39] TechCrunch, "Sean Parker is rebuilding Stability AI around music," 2 October 2026. <https://techcrunch.com/2026/10/02/sean-parker-is-rebuilding-stability-ai-around-music/>.

[40] The Associated Press, "AP, Open AI agree to share select news content and technology in new collaboration," 13 July 2023. <https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/>.

[41] BNN Bloomberg, "Meta strikes multiple AI deals with news publishers," 5 December 2025. <https://www.bnnbloomberg.ca/business/2025/12/05/meta-strikes-multiple-ai-deals-with-news-publishers/>.

[42] Google, The Keyword, "400 publications are now partnering with us on our News AI pilot program," 15 October 2026. <https://blog.google/products-and-platforms/products/news/news-ai-pilot-update-publishers/>.

[43] eWeek, "AI deal between Amazon, New York Times worth up to \$25M annually," 31 July 2025. <https://www.eweek.com/news/amazon-new-york-times-licensing-deal-ai-training/>.

[44] The Decoder, "Axel Springer and OpenAI license agreement is worth 'tens of millions of euros' per year," 19 December 2023. <https://the-decoder.com/axel-springer-and-openai-license-agreement-is-worth-tens-of-millions-of-euros-per-year/>.

[45] interactive investor, "ii view: confident Informa does AI data deal with Microsoft," 8 May 2024. <https://ii.co.uk/analysis-commentary/ii-view-confident-informa-does-ai-data-deal-microsoft-ii531624>.

[46] Universal Music Group, "Universal Music Group and Udio announce Udio's first strategic agreements for new licensed AI music creation platform," 30 October 2025.

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

[47] Al Jazeera, "Sony, Warner and Universal sign AI music licensing deals with startup Klay," 20 November 2025. <https://www.aljazeera.com/economy/2025/11/20/sony-warner-and-universal-sign-ai-music-licensing-deals-with-startup-klay>.

[48] WinBuzzer, "AI training: Microsoft launches Publisher Content Marketplace for content licensing," 4 February 2026. <https://winbuzzer.com/2026/02/04/microsoft-publisher-content-marketplace-ai-licensing-xcxwbn/>.

[49] TechCrunch, "Google is testing AI-powered article overviews on select publications' Google News pages," 10 December 2025. <https://techcrunch.com/2025/12/10/google-is-testing-ai-powered-article-overviews-on-select-publications-google-news-pages/>.

[50] TechCrunch, "Mistral signs deal with AFP to offer up-to-date answers in Le Chat," 16 January 2025. <https://techcrunch.com/2025/01/16/mistral-signs-deal-with-afp-to-offer-up-to-date-answers-in-le-chat/>.

[51] TechCrunch, "OpenAI is bankrolling Axios' expansion into four new markets," 15 January 2025. <https://techcrunch.com/2025/01/15/openai-is-bankrolling-axios-expansion-into-four-new-markets/>.

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). AI data licensing deals, July 2023 to October 2026: a public dataset and what it shows. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-deals>