The Hidden Acquisitions Index: a source-checked dataset of AI licence-and-hire, team-hire, minority-stake and compute-for-equity transactions, 2023–2026

Joash Boyton, Acquiry Labs, Melbourne, Australia · ORCID 0009-0000-2405-8766

Data descriptor · dataset version 1.1.0 (2026-11-19) · https://doi.org/10.5281/zenodo.24244788

Abstract

Large technology companies have gained the staff, technology or influence of AI start-ups through licences, hires, minority stakes and investments tied to cloud commitments, instead of buying the companies. Merger statistics, which count completed purchases, miss most of these transactions. The Hidden Acquisitions Index records 109 such transactions announced between 23 January 2023 and 30 October 2026: 36 core deals in which the buyer gained a control lever, 63 related investments with a commercial tie, and 10 deals held back for incomplete evidence. Every coded fact is linked to one of 1,700 sources, each with an exact quote of at most 20 words. The dataset also records 82 regulator events, 62 disclosed amounts kept as separate components, outcomes for target companies at 6, 12 and 24 months, and 168 candidates excluded with reasons. It is released under CC BY 4.0 with a codebook, a datasheet, and Croissant and Frictionless metadata.

Background & Summary

Since 2023 a recurring transaction pattern has appeared in the AI industry. A large company licenses a start-up’s technology, hires its founders and most of its staff, and pays a fee that the start-up uses to return money to its investors. The start-up continues to exist but often with few staff. Related structures include hiring a team without a licence, taking a minority stake that carries board, observer, veto or consultation rights, and investing in a start-up that commits to spend on the investor’s cloud or chips.

Competition authorities in the United Kingdom, the European Union, the United States, Japan and Australia have examined several of these arrangements. Some reviews turned on whether a licence-and-hire deal amounted to a merger, and others on whether a minority stake conferred material influence. Commentators have used informal names for the pattern, including “reverse acqui-hire”. Researchers who want to study it face practical obstacles: the deals are scattered across press releases, regulator registers, filings and news reports; terms are often undisclosed; and amounts are reported in ways that invite double counting, such as a licence fee, an investment and a cloud commitment announced together.

Other public trackers of these deals exist, including those maintained by SOMO and GenAI Brief. The Hidden Acquisitions Index adds four features intended for reuse in research:

1. **Field-level evidence.** Every coded fact is linked to a source record with an exact quote, the fields it supports, and, where the site allows it, an archived copy.
2. **Money as components.** Each disclosed amount is a separate record with its type and is never summed.
3. **Outcomes at fixed intervals.** What happened to the target is observed at 6, 12 and 24 months.
4. **Published exclusions.** Every candidate that was considered and left out is listed with a reason code and a source.

Intended uses include descriptive research on deal structures and their evolution, studies of regulatory responses and disclosure, teaching, and journalism. The dataset takes no position on whether these transactions are good or bad for competition.

Methods

Scope

A transaction is in scope if a buyer gained an AI company’s people, technology or influence without acquiring control of the company through a purchase of its shares or assets. Five structures are coded:

Code	Structure	What the buyer receives
LICENCE_AND_HIRE	Licence and hire	A licence to the target’s technology plus the hire of founders or staff
TEAM_ONLY_HIRE	Team hire	Founders or staff, with no disclosed licence
MINORITY_STAKE_CONTROL	Minority stake with control rights	A minority stake with a board seat, observer seat, veto or consultation rights
COMPUTE_FOR_EQUITY	Compute for equity	Equity linked to the target’s commitment to use the buyer’s cloud or chips
OTHER_HYBRID	Investment with a commercial tie	A minority investment combined with a commercial partnership

Each deal is assigned to a tier. **Core** deals give the buyer at least one control lever: staff, a licence, a board seat or a veto. **Related** deals are investments with a commercial tie but no recorded control lever. **On-hold** deals have incomplete evidence. They are published with that status, but they are excluded from all counts on the website and from the descriptive statistics below. The rule applied to each deal is stored in the `tier_rule` and `tier_reason` fields.

Identifying candidates

Candidates came from four streams:

- regulator registers and decisions, including notifications to the European Commission under Article 14 of the Digital Markets Act and UK Competition and Markets Authority merger inquiries;
- company announcements and securities filings;
- court and bankruptcy records;
- reporting by major newsrooms.

The coverage window runs from the first in-scope announcement found (23 January 2023) to 30 October 2026.

Inclusion and exclusion

Each candidate was tested against written rules. Candidates that failed are kept in the exclusions table with a reason code:

Code	Reason	Candidates
E1	Full acquisition	109
E2	No control rights	14
E3	Not an AI startup	7
E4	Insufficient evidence	31
E6	Other	7

Source standard and grading

A fact may be coded only from an eligible source.

- **Grade A** sources are the company itself, a securities filing, a regulator, a court or a company registry.
- **Grade B** sources are major newsrooms with editorial standards. A grade B fact requires two independent newsrooms, where one does not simply cite the other.

Wikipedia, commercial deal databases, blogs, social media posts and aggregator sites are used only to find original sources and are never cited.

Each source record holds the publisher, title, date, grade, URL, an exact quote of at most 20 words, the fields the quote supports, an archived copy where available, and, where relevant, the source it derives from. Of the 1,700 sources, 1,460 are grade A and 240 are grade B.

Coding

The codebook defines 110 fields, each with a type, allowed values, a definition and a coding rule. Four conventions matter most for reuse:

- **“Not disclosed”** means that no eligible source states the value. It is neither zero nor “no”.
- **Conflicts** are preserved. Where eligible sources disagree, each version is kept with its source, and the disagreement is described in the `conflict` field.
- **Money** is recorded exactly as stated, with the source’s own qualifier (“up to”, “agreed to invest”, “valued at”), and classified as paid, committed or ceiling, invested, conditional, or a spending commitment by the target. Amounts are never converted, estimated or summed.
- **Dates** are ISO 8601. A precision field records whether a date is known to the day, month or year. Where no announcement exists, the date is the first substantiated public report or the date of a regulatory notification.

Outcomes

For each deal, the status of the target and of the business left behind is recorded with the date it was last confirmed. It is also observed at 6, 12 and 24 months after announcement where that interval has passed.

Data Records

The dataset is deposited on Zenodo (https://doi.org/10.5281/zenodo.24244788) and mirrored at https://www.acquiry.com/labs/hidden-acquisitions/. Every table is provided as CSV and as JSON.

File	Rows	Content
deals	109	One row per deal: parties, dates, structure, rights, values, outcomes, tier

sources	1,700	One row per source: publisher, title, date, grade, URL, archived copy, exact quote, fields supported, check record
regulatory_events	82	Dated actions by competition authorities and other public bodies, with URL
money_components	62	Each disclosed amount as stated, with type, currency and quote
rejected	168	Candidates excluded, with reason code, explanation and source
codebook	110	Every field: type, allowed values, definition, coding rule

Machine-readable metadata are supplied as a Croissant file (MLCommons) with SHA-256 checksums, a Frictionless Data package with field schemas and foreign keys, and a CITATION.cff file. A datasheet following Gebru et al. answers the standard questions on motivation, composition, collection, uses and maintenance. Versions are frozen on Zenodo, and each version has a changelog entry.

Technical Validation

Exact-quote checks. On 18 November 2026 a deterministic script (verify_sources.py, run on GitHub Actions) fetched every one of the 1,700 sources, or its archived copy, and searched the extracted text for the stored quote after normalising whitespace, quotation marks and dashes. It found 1,385 exact matches, 45 near matches (95% similar or more), 119 pages where the quote was not found and 150 pages that refused automated access or failed to load. Most “not found” results came from the European Commission’s Digital Markets Act register, which shows its text only after running JavaScript; all 82 of those quotes were confirmed on the page as rendered in a browser. The 412 sources added in research round 6 were also checked separately, before merging, for whether the quote appears on the page, whether it supports the coded value and whether the date is right. Each source record keeps the result of every check, so users can restrict analysis to automatically confirmed sources.

Archiving. 1,635 sources (96%) have an archived copy in the Internet Archive’s Wayback Machine. The rest come from sites that block or limit archiving (for example some newsrooms and company registers); each such record says so.

Independent re-research. For version 1.1.0, every blank field in every live deal was researched again from scratch under the same source standard. Every new or changed value was then checked against its quoted source before it was merged.

Cross-file consistency. An automated audit, run before each release, confirms that: - counts agree across the data files, the metadata files, this paper and the website; - checksums match; - every column is defined in the codebook; - every quote is 20 words or fewer; - no amounts are summed in any text; - the documentation makes no claim that the data do not support.

Known limitations of validation. Classification into structures and tiers involves judgement. The written rules, the per-deal tier reason and the exclusions table are published so that users can apply different rules. Version 1.1.0 does not report an inter-coder reliability statistic.

Usage Notes

Coverage overview. Counts here exclude the 10 on-hold deals, leaving 99 deals that involve 31 buyers and 69 target companies. - The core deals comprise 19 licence-and-hire deals, 13 team hires and 4 minority stakes with control rights. - The related deals comprise 47 investments with a commercial tie, 15 compute-for-equity investments and 1 minority stake with control rights. - The most frequent buyers are NVIDIA (15), Google (11), SK Telecom (9), Microsoft (8), Amazon (6), Apple (6). - The most common target countries are United States (57), Japan (11), Canada (6).

Year	All live deals	Core	Licence and hire	Team hire	Minority stake
2023	17	2	0	0	2
2024	21	6	5	0	1
2025	26	11	5	5	1
2026 (to 30 October)	35	17	9	8	0

Regulators. 71 actions by six competition authorities relate to 24 deals: European Commission (27); UK Competition and Markets Authority (22); Australian Competition and Consumer Commission (9); US Federal Trade Commission (6); Japan Fair Trade Commission (5); Bundeskartellamt (2). Users should note that notifications under Article 14 of the Digital Markets Act are made by the buyer, not initiated by the authority, and often precede public reporting of the deal. A further 10 events involve other public bodies, such as letters from legislators and court filings.

Outcomes for core targets. Recorded outcomes are 11 continued, 5 product discontinued, 4 wound down, 3 bankruptcy, 1 pivoted, 1 sold later, 1 sold assets. For 10, the status after the deal is not disclosed.

Disclosure. Terms are often private. Among the 99 live deals, a disclosed amount is recorded for 30 and an equity percentage for 12. The number of staff who moved is recorded for 16, a voting board seat for 10 and an exclusivity term for 18. Users should treat these rates as findings about disclosure, not as the prevalence of the underlying terms.

Do not: - sum money components, within a deal or across deals; - read “not disclosed” as “no”; - infer legality or competitive effects from the coding alone.

Corrections and updates. Corrections can be reported through https://www.acquiry.com/contact and are logged in the changelog. New versions are deposited under the same concept DOI.

Code Availability

The scripts that check quotes, build the release files, compute the checksums and run the cross-file audit are deposited with the dataset.

Competing Interests

The author is the editor of the dataset. He works for Acquiry Pty Ltd, which publishes it and which advises on acquisitions. No company pays to be included or excluded, and deals Acquiry has advised on are excluded. No external funding was received.

References

Akhtar, M. et al. Croissant: a metadata format for ML-ready datasets. *Proc. Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM ’24)*, 1–6 (2024).

Gebru, T. et al. Datasheets for datasets. *Commun. ACM* **64**(12), 86–92 (2021).

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector (Digital Markets Act). *OJ L* **265**, 1–66 (2022).