Valuing Proprietary Training Data in AI M&A

A Transaction Framework for Rights Uniqueness Refresh Cost and Model Contribution

Authored by Chennakeshav Adya
Independent Researcher
September 2026

Abstract

Proprietary training data can support differentiated artificial-intelligence products, yet the word proprietary often conceals several different assets. A company may own a database while holding only limited rights to train a model. It may possess lawful access without exclusivity. It may control labels created by employees while relying on source material licensed from customers or third parties. A dataset may be expensive to reproduce and still contribute little incremental performance. Another dataset may be small, current and operationally embedded, creating substantial value through faster decisions, lower error rates or access to a scarce workflow.

This paper develops a transaction framework for valuing training data in AI mergers and acquisitions. It separates legal control, provenance, technical quality, uniqueness, refresh economics, model contribution, workflow adoption and cash realisation. The framework treats training data as a governed production system rather than a static file. It links each valuation conclusion to evidence that a buyer can inspect, reproduce and protect after closing.

The analysis draws on intellectual-property and valuation guidance from the World Intellectual Property Organization; accounting concepts in IAS 38, IFRS 3 and IFRS 13; privacy guidance from the European Data Protection Board and the United Kingdom Information Commissioner's Office; the European Union's AI Act, Data Act, General Data Protection Regulation, Trade Secrets Directive and Database Directive; United States Copyright Office work on generative-AI training; NIST AI-risk and provenance guidance; OECD principles; and recognised data, security and AI-management standards [1-50]. These sources establish useful legal, accounting, governance and technical reference points. They do not determine the rights, lawful basis, performance, accounting treatment or value of a specific asset.

An illustrative acquisition shows the method. A target reports USD 210 million of revenue and USD 34 million of EBITDA. Management attributes USD 110 million of enterprise value to its training-data estate. The framework identifies a USD 42 million replacement-cost indication, a USD 76 million income indication and a USD 58 million evidence-weighted conclusion after deductions for uncertain source rights, customer restrictions, label decay, refresh obligations and dependence on the target's workflow. Every amount is a management assumption used only to demonstrate the method.

The central conclusion is that a buyer should pay for enforceable, reproducible and renewable economic advantage. Dataset size, acquisition cost and model performance in a historical benchmark are supporting facts rather than standalone measures of value. The most defensible conclusion reconciles cost, income and market evidence; isolates the dataset's incremental contribution; reflects legal and operational constraints; and places unresolved claims into price adjustments, escrow, indemnities, earn-outs or staged access arrangements.

JEL Classification: G24, G34, K11, K24, L86, M41, O32, O34

Keywords: training data, artificial intelligence, data valuation, mergers and acquisitions, intellectual property, data rights, model contribution, intangible assets

Introduction

Data is routinely described as an asset in AI transactions. That description is directionally useful and financially incomplete. A buyer cannot value a training-data estate until it knows what the target controls, what uses are permitted, which technical characteristics affect model performance, how quickly the data decays and which cash flows depend on continued access. A folder containing images, conversations or sensor histories is not equivalent to an enforceable right to develop and commercialise a model.

The issue has become more important as AI companies seek capital, strategic partnerships and exits. Buyers may encounter claims based on record count, annotation expenditure, years of collection, rarity, domain coverage or the performance of a model trained on the dataset. Each claim can be relevant. Each can also overstate value when provenance is incomplete, rights are narrow, labels are inconsistent, records are duplicated, customers can revoke access, the data is stale or the model would perform nearly as well without it.

WIPO identifies data, algorithms and trade secrets as emerging intangible assets that present valuation challenges, and its IP-valuation materials emphasise identifiability, evidence, transferability and measurable economic benefit [1-4]. Financial-reporting standards distinguish recognition from economic value and require acquisition-date analysis of identifiable intangibles in business combinations [5-8]. Privacy, copyright, database, trade-secret, competition and AI rules can shape permitted use and transfer [9-26]. Technical frameworks emphasise provenance, documentation, testing, security and lifecycle management [27-40].

This paper is designed for boards, corporate-development teams, private-equity investors, venture investors, lenders, founders, valuation specialists and integration leaders. It provides a transaction decision system. It does not provide legal, regulatory, accounting, tax, technical-security or valuation advice. Applicable law, contract terms, technical evidence and market circumstances require target-specific professional review.

1 Define the asset before discussing value

The phrase training data may refer to raw source records, cleaned observations, labels, features, prompts, preference rankings, synthetic records, evaluation sets, red-team cases, human feedback, retrieval corpora or the data pipeline that continuously produces them. These components can have different owners, restrictions, economic lives and contributions. A single headline valuation therefore needs a component inventory.

The inventory should identify the record class, originating party, collection event, applicable contract, lawful basis, permitted purpose, territory, retention rule, transformation history, quality controls, security classification, model use and commercial workflow. It should distinguish records that the target owns from records it licenses, hosts, accesses under customer instructions or obtains from public sources. It should also distinguish data rights from software, models, know-how, customer relationships and assembled workforce.

A useful unit of account is the smallest collection that has a coherent rights package and measurable economic use. A clinical image library licensed for one diagnostic task may be one unit. Technician-labelled failure histories for a defined equipment family may be another. Customer conversations gathered under different terms should be separated by contract cohort. Combining them into one data lake does not combine the rights.

The buyer should map dependencies. Labels may rely on specialist annotators, customer confirmations or later outcomes. Features may depend on proprietary software. A dataset may be valuable only with a taxonomy, ontology, retrieval layer or feedback loop. The transaction perimeter should show whether each dependency transfers, remains available under a transition arrangement or needs replacement.

Table 1 Training data asset inventory

Asset component	Evidence of control	Economic use	Principal valuation question
raw source records	contracts consents notices collection logs and access controls	observation of the underlying domain	can the buyer lawfully retain transfer and reuse the records
labels and annotations	employment supplier or customer terms plus quality records	supervised learning and evaluation	who owns the labels and how reproducible is their quality
derived features	transformation code lineage and documentation	improved model efficiency or accuracy	are the features separately transferable or inseparable from software
evaluation sets	locked versions sampling rules and test governance	evidence of generalisation and safety	are results independent representative and free from contamination
feedback and preference data	collection terms reviewer instructions and audit trail	alignment ranking and product improvement	does the target control continued collection after closing
retrieval corpus	source licence index history and update process	grounded responses and workflow knowledge	what content can be retrieved reproduced and commercialised
data production system	people contracts tools controls and refresh schedule	continued creation of current data	what recurring cost and operational dependency sustains the asset

Proposed diligence record; legal accounting tax privacy competition and technical specialists should test each material data class.

2 Establish the rights chain and transaction perimeter

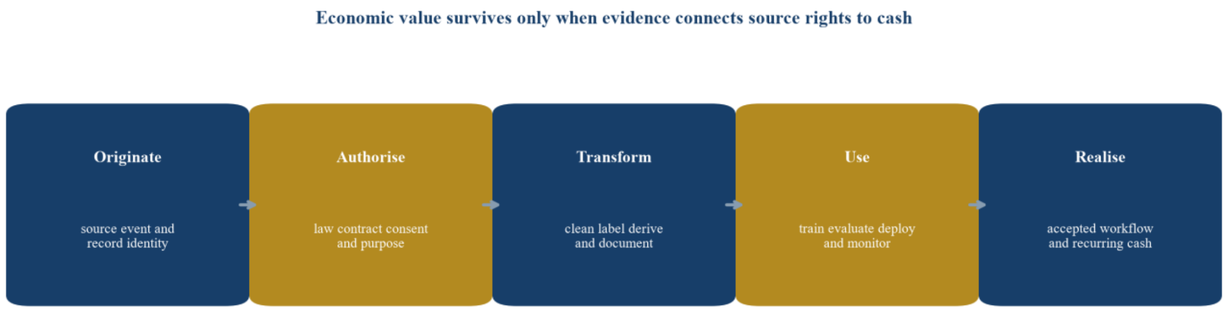
Value begins with enforceable use. WIPO's valuation guidance links quantifiable value to identifiable assets, evidence of existence, enforceability, transferability and measurable benefit [2-4]. Training data often has overlapping rights and duties. Copyright may protect source works or selection and arrangement. Database rights may apply in relevant jurisdictions. Trade-secret protection may depend on secrecy and reasonable protective measures. Contracts can create broader or narrower permissions. Privacy law can regulate processing even when the dataset is commercially controlled.

The buyer should reconstruct the rights chain from origin to closing. For each material source, it should identify who created or collected the information, which notices and permissions applied, which entity contracted with the source, what transformations occurred, where the data and model are hosted, who can access them, what customer or contributor rights remain and whether change of control affects use. The analysis should cover training, fine-tuning, evaluation, inference, retrieval, benchmarking, product improvement, sublicensing and post-termination retention separately.

Permission to provide a service does not automatically establish permission to train a general model. A licence to use content may exclude machine learning, derivative datasets or redistribution. A customer agreement may authorise improvement of that customer's instance while prohibiting pooled learning. Public availability does not itself resolve copyright, privacy, contract, confidentiality or database questions. The United States Copyright Office's work on generative-AI training and the European Data Protection Board's opinion on AI models illustrate the continuing importance of source-specific and use-specific analysis [15-18].

Rights should be scored by evidence maturity. A signed agreement with an explicit training and transfer clause carries more weight than a historic policy, a general representation or an undocumented practice. The score should reflect revocability, exclusivity, term, territory, field of use, sublicensing, audit rights, indemnity, survival and change-of-control provisions. A dataset with narrow but clear rights may be more valuable than a larger collection whose use remains disputed.

Figure 1 Rights to cash evidence chain



Proposed transaction map; target-specific counsel should determine the effect of applicable law contracts notices and technical controls.

3 Measure uniqueness without confusing scarcity and usefulness

Uniqueness has several dimensions. A dataset may contain observations that competitors cannot readily obtain. It may cover rare events, difficult environments, specialist judgments or outcomes observed over a long period. It may have unusually consistent labels, rich context or a direct link between intervention and result. These features can support value when they improve a commercial decision.

Scarcity alone is insufficient. A rare dataset can be irrelevant to the buyer's target task. A large corpus can be redundant because similar information is licensed cheaply or generated through ordinary operations. A proprietary collection can lose distinctiveness when public datasets, synthetic data, customer-owned data or improved foundation models reduce the marginal benefit of the original records.

The buyer should compare the target dataset against realistic alternatives. The comparison should include direct licensing, renewed collection, partnerships, customer-contributed data, public sources, simulation, synthetic augmentation and acquisition of another company. It should consider time to obtain rights, time to achieve usable quality, domain coverage, rare-event density, annotation reliability, refreshability and the buyer's ability to integrate the alternative.

Uniqueness should be tested at the decision level. For a fraud model, rare confirmed outcomes and changing attack patterns may matter more than gross transaction count. For industrial maintenance, failure histories linked to operating conditions and interventions may matter more than unlabelled sensor volume. For an enterprise assistant, current permissions and retrieval quality may matter more than the age of the corpus. For autonomous systems, edge cases and scenario coverage may matter more than average-condition observations.

The target should provide nearest-alternative evidence. Useful records include data-licensing quotes, internal collection budgets, annotation yields, duplicate analysis, overlap tests, domain coverage, benchmark results and expert interviews. The buyer should test whether management selected comparisons that make the dataset look artificially scarce.

Table 2 Uniqueness and substitutability test

Dimension	Strong evidence	Weak evidence	Valuation effect
source exclusivity	enforceable exclusive or structurally privileged access	internal assertion of ownership	supports scarcity only for the permitted use and term
rare-event coverage	verified event counts with outcomes and representative sampling	gross record volume	may reduce collection time and model error in valuable cases
label quality	independent agreement adjudication and outcome linkage	single untested annotator	affects usable quantity and retraining cost
contextual depth	linked operating conditions interventions and results	isolated records without context	can improve causal interpretation and workflow fit
alternative availability	current quotes and tested substitutes show material delay or loss	no documented market search	supports replacement-time and income advantage
refresh access	renewable sources and continuing rights	one-off historic snapshot	extends economic life and supports continuing differentiation

Proposed scoring framework; conclusions should be based on current target and alternative evidence.

4 Reconstruct provenance quality and usable quantity

Record count is a starting point. Usable quantity is lower after removing duplicates, corrupted files, records outside the permitted purpose, unresolved identities, inconsistent labels, contaminated evaluation examples and observations that do not represent the intended deployment population. NIST's generative-AI profile emphasises training-data provenance and documentation as risk-management practices [27-30]. ISO and OECD materials similarly support systematic governance, traceability and accountability [31-36].

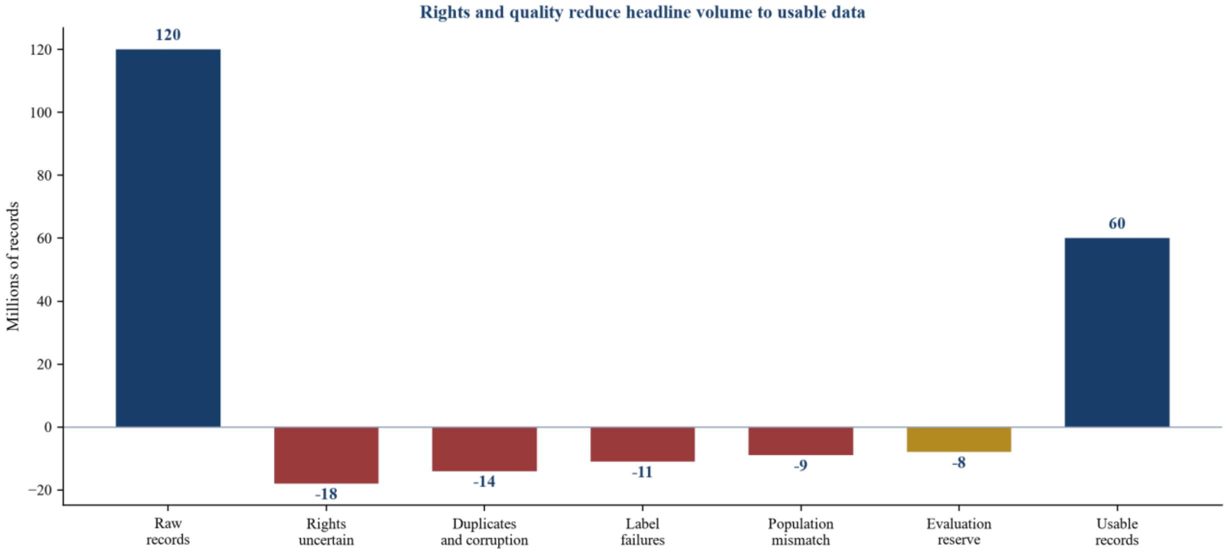
The buyer should reproduce the target's data lineage. It should select representative records from each source cohort and trace them through collection, cleaning, annotation, transformation, feature generation, training, evaluation and deletion processes. Hashes, version identifiers, manifests, access logs and code commits should support the trace. A polished data catalogue without record-level correspondence is weaker evidence.

Quality should be defined by the task. Relevant dimensions can include accuracy, completeness, consistency, timeliness, coverage, class balance, label agreement, outcome maturity, measurement error and robustness to distribution shift. The buyer should test whether quality metrics were calculated across the whole estate or only on a curated subset. It should identify records excluded from production and the reasons for exclusion.

Evaluation leakage deserves specific attention. If examples from the same customer, patient, asset, document family or time period appear in both training and test sets, reported performance can overstate generalisation. Benchmark contamination can produce a similar effect. The buyer should inspect split logic, near-duplicate detection, temporal holdouts and independent replication. Locked evaluation sets need access controls and version governance.

Usable quantity should be reported as a waterfall. Starting records are reduced for rights uncertainty, duplication, corruption, inadequate labels, population mismatch and evaluation isolation. The remaining data should be grouped by task, geography, period and customer cohort. This provides a more useful valuation input than a single terabyte or record figure.

Figure 2 Illustrative usable data waterfall



Millions of records; management assumptions used only to demonstrate the method.

5 Isolate the dataset's contribution to model and workflow performance

Training data creates economic value through an effect on a product or decision. The target should demonstrate that effect with controlled comparisons. A model trained with the claimed dataset should be compared with a credible baseline trained without it, with smaller samples, with alternative sources and with the buyer's existing data. The test design should hold architecture, compute, tuning and evaluation conditions sufficiently constant to isolate contribution.

Performance should be measured in the context of use. Accuracy, precision, recall, calibration, ranking quality, latency, robustness and hallucination rates can matter. The economic metric may be avoided loss, lower review time, higher conversion, reduced downtime, faster approval or improved retention. A statistically visible improvement may have little cash value if it occurs on low-value cases. A modest average improvement may be valuable if concentrated on rare and costly events.

The buyer should require confidence intervals, subgroup results and temporal tests. It should identify where the dataset improves performance, where it has no effect and where it harms performance. The target should disclose failed experiments and selection decisions. Results generated by the seller should be independently reproduced on a controlled environment when practical.

Contribution can interact with model architecture and workflow design. A dataset may be valuable with one representation and less useful after a foundation-model upgrade. A retrieval corpus may create value through factual grounding without changing model weights. Human feedback may matter because it reflects a specific customer workflow. The valuation should specify the configuration in which contribution was observed and the cost of maintaining it.

The buyer should also test model extraction and memorisation risk. If protected or personal information can be elicited from a model, the economic asset may carry remediation, claims and deployment restrictions. EDPB guidance treats anonymity and unlawful processing as case-specific questions with potential consequences for later deployment [16-18]. Security testing should therefore connect data lineage to model behaviour.

Table 3 Incremental contribution test

Comparison	What it isolates	Minimum evidence	Common failure
target data versus no target data	gross incremental effect	fixed architecture compute tuning and evaluation set	other changes are attributed to data
full dataset versus size cohorts	diminishing returns	learning curves and repeated samples	largest run is compared with an under-tuned baseline
target data versus alternative source	substitutability	equivalent rights quality and domain tests	weak alternative is selected
historic versus current cohort	decay and drift	time-based holdout and refresh record	old benchmark masks current deterioration
model metric versus workflow outcome	economic translation	accepted operational measure and accountable owner	technical gain lacks adoption or cash evidence
central versus subgroup results	representativeness and harm	material customer geography and population cohorts	average result hides weak or harmful segments

Proposed evidence record; test design should be independently reviewed for each material use case.

6 Model refresh cost decay and economic life

Training data is often a wasting and renewable asset at the same time. Labels can become stale, product taxonomies change, language evolves, fraud tactics adapt, sensors are replaced and customer behaviour shifts. The dataset may require continuous acquisition, adjudication, deletion, security, documentation and revalidation. These recurring activities affect both sustainable earnings and the asset's economic life.

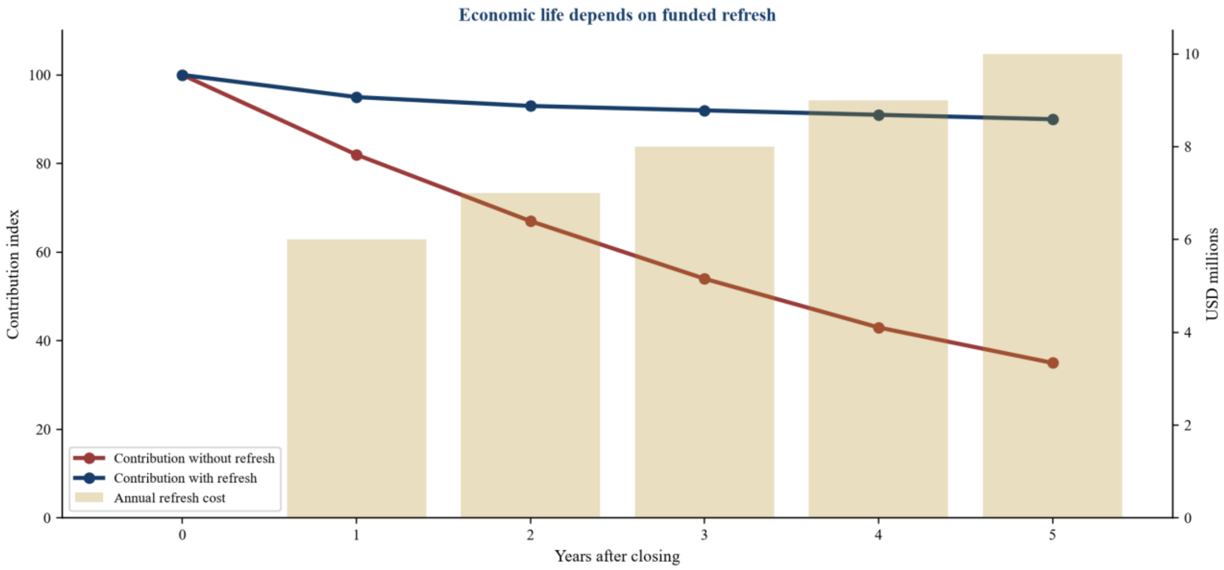
The buyer should create a refresh ledger for every material data class. The ledger should show observation frequency, collection yield, consent or contract renewal, annotation hours, specialist review, dispute resolution, quality-control sampling, storage, security, privacy operations, model retraining, validation and retirement. It should identify which activities are performed by scarce employees, customers, contractors or embedded product workflows.

Historical expenditure should be separated into creation, maintenance and failed work. Replacement cost should exclude waste that a rational buyer would avoid. It should include the current cost of obtaining comparable lawful rights, rebuilding taxonomies, recruiting specialists, reproducing labels, collecting rare outcomes, establishing controls and waiting for results to mature. Time can create a significant opportunity cost where outcomes require years of observation.

Economic life should be linked to the period over which the dataset provides incremental cash before replacement or substantial refresh. Contract terms, legal changes, customer churn, model architecture, source continuity, competition and domain drift can shorten that period. A perpetual growth assumption is rarely compatible with a finite rights term and high refresh dependence.

Refresh cost also affects EBITDA. If management capitalises data work or classifies essential annotation and governance as growth investment, reported operating profit may overstate sustainable earnings. The buyer should normalise the cost required to maintain current revenue and performance before applying a transaction multiple.

Figure 3 Illustrative data contribution and refresh profile



Index values and USD millions; management assumptions used only to demonstrate the method.

7 Reconstruct sustainable earnings and avoid double counting

A training-data valuation should be consistent with the enterprise valuation. If forecast revenue and margins already reflect the product advantage created by proprietary data, adding the full value of that advantage as a separate asset can double count it. The valuation bridge should identify where each benefit enters the model and how contributory assets are charged.

The illustrative target reports USD 210 million of revenue and USD 34 million of EBITDA. Diligence identifies USD 5 million of genuine non-recurring expense. It also identifies USD 8 million of recurring annotation and data acquisition, USD 4 million of privacy, provenance and security operations, and USD 3 million of model evaluation and customer-specific maintenance that are necessary to sustain current performance but are incompletely reflected. Sustainable EBITDA is therefore USD 24 million. Every figure is a management assumption used only to demonstrate the method.

The income attributable to data is not total product income. Software, model architecture, compute, brand, customer relationships, distribution, workforce and working capital also contribute. A multi-period excess-earnings analysis should apply charges for these contributory assets. A with-and-without analysis should model the cash-flow difference between continued lawful access to the data and the best realistic alternative. A relief-from-royalty method may be considered where comparable licensing evidence exists and the licensed asset can be defined consistently.

The buyer should reconcile the data valuation to purchase-price allocation without assuming that accounting recognition equals deal value. IAS 38 defines identifiable intangibles through separability or contractual and legal rights, while IFRS 3 and IFRS 13 govern relevant business-combination and fair-value concepts [5-8]. Professional accounting and valuation judgment is required for the actual transaction.

Table 4 Illustrative sustainable EBITDA bridge

Item	Amount	Treatment	Reason
reported EBITDA	34	starting point	seller-reported operating result
genuine non-recurring expense	5	add	separately evidenced one-time cost

Item	Amount	Treatment	Reason
recurring annotation and acquisition	8	deduct	required to maintain current data coverage
provenance privacy and security operations	4	deduct	recurring control needed for lawful trusted use
model evaluation and customer maintenance	3	deduct	recurring cost needed to sustain acceptance
sustainable EBITDA	24	conclusion	illustrative maintainable operating result

USD millions; management assumptions used only to demonstrate the method.

8 Apply three valuation approaches and reconcile them

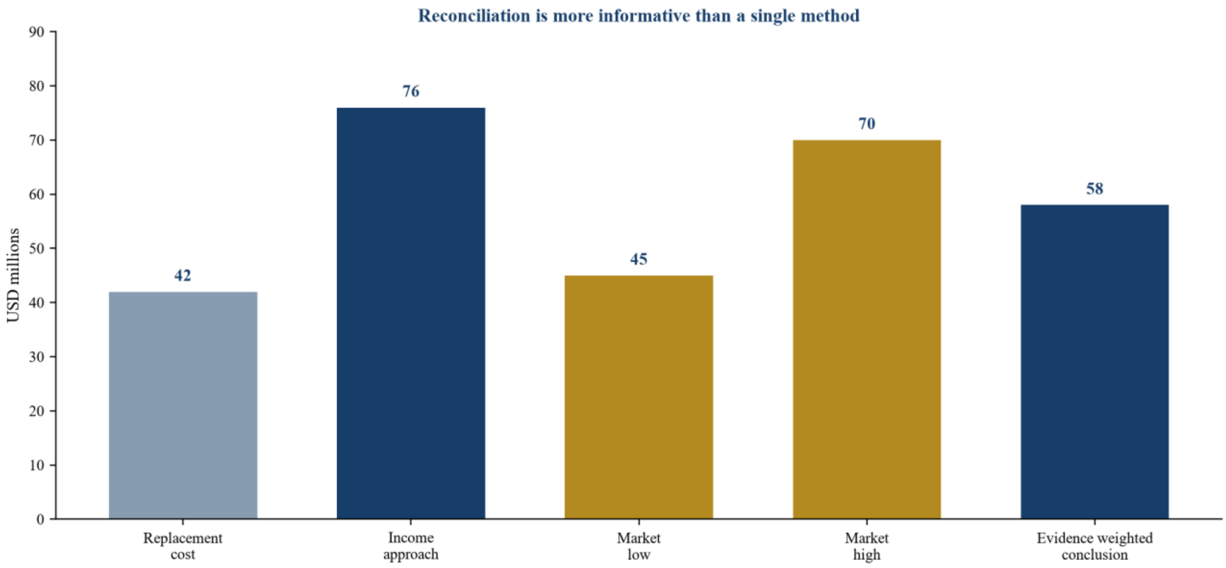
The cost approach estimates the current expenditure and time required to recreate an asset of equivalent utility. It is useful where the buyer can define the replacement specification and observe collection, licensing, annotation, control and validation costs. Cost does not capture all economic benefit. It can overvalue wasted effort and undervalue rare observations, time advantage, exclusive access or a proven feedback loop.

The income approach estimates cash flows attributable to the data and discounts them for risk. A with-and-without method compares the business with continued access against the best available alternative. A multi-period excess-earnings method deducts contributory-asset charges. A relief-from-royalty method estimates avoided licence payments where defensible comparable evidence exists. Income methods require careful separation of data contribution from model, software, workforce, customer and brand effects.

The market approach uses prices or licence terms for comparable assets. Training datasets are heterogeneous and transaction details are often confidential. Comparability should cover rights, exclusivity, domain, quality, scale, rare-event density, recency, permitted use, geographic scope, refresh access and buyer-specific synergies. A price per record can mislead when records differ in utility and rights.

The illustrative replacement-cost analysis begins with USD 51 million of current collection, annotation, validation, governance and waiting-time cost. It deducts USD 9 million for avoidable historic inefficiency, producing USD 42 million. The income analysis produces USD 76 million after contributory-asset charges, refresh cost, tax and risk. A limited market analysis indicates USD 45 million to USD 70 million. The evidence-weighted conclusion is USD 58 million after considering rights uncertainty, concentration and refresh dependence. These figures demonstrate the method and do not describe a company or market.

Figure 4 Illustrative training data valuation indications



USD millions; management assumptions used only to demonstrate the method.

9 Convert uncertainty into an evidence weighted conclusion

Valuation uncertainty should be expressed through scenarios, ranges and explicit deductions rather than a false point estimate. The buyer should distinguish factual uncertainty from commercial optionality. Missing source contracts, unresolved deletion duties or untested model contribution reduce confidence in the existing asset. A possible future licensing product is an option that requires investment and market evidence.

An evidence-weighting model can score rights, provenance, uniqueness, quality, contribution, refreshability, security and cash realisation. Weighting should reflect the transaction thesis. A buyer acquiring an AI safety platform may place greater weight on representative evaluation data and incident outcomes. A buyer acquiring a vertical workflow product may emphasise customer permissions, outcome linkage and embedded refresh.

Scores should change the valuation mechanics. Rights uncertainty can reduce the eligible dataset, shorten life, increase probability-weighted remediation cost or support a specific indemnity. Weak contribution evidence can move value from upfront consideration to an earn-out. Customer concentration can affect both forecast cash and continued data access. High refresh dependence can reduce sustainable margin and require a funded operating plan.

The conclusion should reconcile with the overall enterprise value. In the illustrative case, management's USD 110 million claim exceeds each independently supported indication. The evidence-weighted USD 58 million amount is incorporated within the enterprise valuation rather than added mechanically. A separate USD 12 million reserve is assumed for remediation, relicensing and deletion work. The buyer should avoid both double counting and the opposite error of ignoring a data asset already reflected in earnings.

Sensitivity analysis should expose the assumptions that move value. The most important variables will often be the percentage of records with durable transfer rights, the remaining rights term, the rate at which model contribution decays, the annual refresh cost, the time needed to build an alternative, customer retention and the portion of workflow cash that can reasonably be attributed to data. The board should see changes to each variable separately and in coherent downside combinations. Correlations matter because loss of a major customer can reduce revenue, remove a refresh source and narrow the deployment population at the same time.

The valuation date should also be explicit. Rights, records, models and alternatives can change quickly. A later foundation-model release, regulatory decision, customer amendment or newly available dataset can alter the conclusion. The transaction model should therefore include a closing-date update process. Material changes between signing and closing should be routed to the responsible workstream, incorporated into the eligible-data register and reflected in price, conditions or the integration plan where appropriate.

Table 5 Illustrative evidence weighted data value

Factor	Evidence assessment	Valuation treatment	Illustrative effect
replacement specification and cost	strong	supports cost floor for equivalent utility	42
attributable income	moderate	discounted for contribution and adoption uncertainty	76
market references	limited	used as a broad reasonableness range	45 to 70
source rights and transfer	moderate	deduct for unresolved cohorts and consent work	negative 7
label recency and refresh	moderate	shorten life and include annual refresh cost	negative 5
workflow adoption	strong in core segment	retain evidence-backed income in that segment	positive support
evidence weighted conclusion	reconciled	embedded within enterprise value	58

USD millions and qualitative scores; management assumptions used only to demonstrate the method.

10 Design diligence around reproducible evidence

Data diligence should be conducted through a controlled environment. The seller can provide a data-room inventory, contract matrix, lineage extracts, representative samples, model cards, evaluation reports, access logs, security evidence and cost records. Highly sensitive records may remain in a clean room or seller-controlled environment. The buyer should agree sampling, queries, permitted outputs and destruction rules before access.

Legal, technical, commercial, finance and security teams should use one asset taxonomy. Separate workstreams often create gaps. Counsel may review source agreements without knowing which data cohorts drive performance. Data scientists may test a dataset without seeing purpose restrictions. Finance may model margin without including refresh cost. A shared identifier for each data class connects these conclusions.

Reperformance should focus on material claims. The buyer can trace selected records, reproduce deduplication, inspect annotation agreement, rebuild time-based splits, rerun ablation tests and reconcile refresh expenditure to the ledger. It can compare the data catalogue with storage and access logs. It can test whether deleted or restricted records persist in derived datasets or models.

The seller should disclose known disputes, takedown requests, licence limitations, security incidents, regulator correspondence and customer objections. It should identify where provenance is inferred from system history rather than supported by source documentation. Representations should follow the evidence and avoid absolute claims that no diligence process can substantiate.

Table 6 Data diligence evidence room

Workstream	Core evidence	Reperformance	Decision output
rights	source contracts notices consents licences and transfer terms	sample records to governing documents	eligible use and transfer matrix
provenance	manifests hashes lineage logs and version history	trace records from source to model input	confidence by data cohort
quality	duplicate error label agreement and coverage reports	rerun selected controls	usable quantity and remediation plan
contribution	ablation learning curve and temporal tests	reproduce material experiments	attributable performance and cash
economics	historic cost headcount supplier invoices and forecasts	rebuild replacement and refresh ledger	sustainable margin and cost indication
security	access control encryption incident and deletion evidence	test selected controls and model leakage	risk reserve and closing conditions
commercial	customer use adoption retention and outcome evidence	reconcile workflow results to contracts and cash	income indication and concentration risk

Proposed work plan; access should follow confidentiality privacy security and competition controls.

11 Translate findings into transaction terms

Price should follow evidence maturity. An asset with clear transferable rights, reproducible contribution and renewable sources can support upfront value. Unresolved rights, untested contribution or customer-dependent refresh may justify deferred consideration. Transaction design can allocate uncertainty without pretending it has disappeared.

Representations can address ownership, licences, permitted use, privacy compliance, trade-secret measures, security, provenance records, disputes, deletion duties and material model uses. Their scope, qualifiers, survival and remedies require transaction counsel. Specific indemnities may address identified cohorts, claims or remediation obligations. Escrow can support recovery where exposure is measurable and time bound.

Earn-outs should use observable outcomes within the buyer's control. Possible measures include retained eligible records, customer renewals that preserve training rights, independently reproduced performance, accepted deployment, recurring revenue or gross profit after refresh cost. Gross dataset size is usually a weak earn-out metric because it can reward low-quality volume.

Carve-outs or staged access may be appropriate where only part of the estate has clear rights. The buyer can exclude restricted data, license a defined corpus, require pre-closing remediation or fund a post-closing rights conversion programme. Transition services may preserve access to annotation teams or data-generation workflows while the buyer establishes its own controls.

Closing conditions can require delivery of a verified data inventory, transfer of key licences, customer consents, deletion of ineligible copies, remediation of critical security gaps and successful reproduction of agreed tests. The board should understand which items affect closing, price, covenants or the integration plan.

Table 7 Transaction protection linked to training data evidence

Finding	Potential structure	Evidence after closing	Release condition
uncertain source rights	cohort exclusion escrow or specific indemnity	executed licence or documented deletion	verified lawful replacement or expiry of claim period
unproven incremental performance	contingent consideration	independent ablation and accepted workflow result	agreed performance and commercial threshold
customer-dependent refresh	retention or consent earn-out	renewed contracts and continuing lawful feeds	defined retained revenue and data rights
weak provenance	remediation covenant and holdback	completed lineage and sample verification	agreed coverage and exception closure
high refresh cost	price adjustment and operating covenant	actual cost and data-quality dashboard	sustainable economics within agreed range
leakage or memorisation risk	security condition and reserve	completed testing remediation and monitoring	independent acceptance against agreed criteria

Illustrative structure; transaction counsel should tailor terms to facts law bargaining position and remedy limitations.

12 Integrate without destroying the asset

Integration can reduce value if the buyer commingles datasets before resolving rights, changes identifiers without preserving lineage, removes specialists who maintain labels or shifts the model to a workflow where the data has not been validated. The integration plan should preserve reversibility until the evidence supports consolidation.

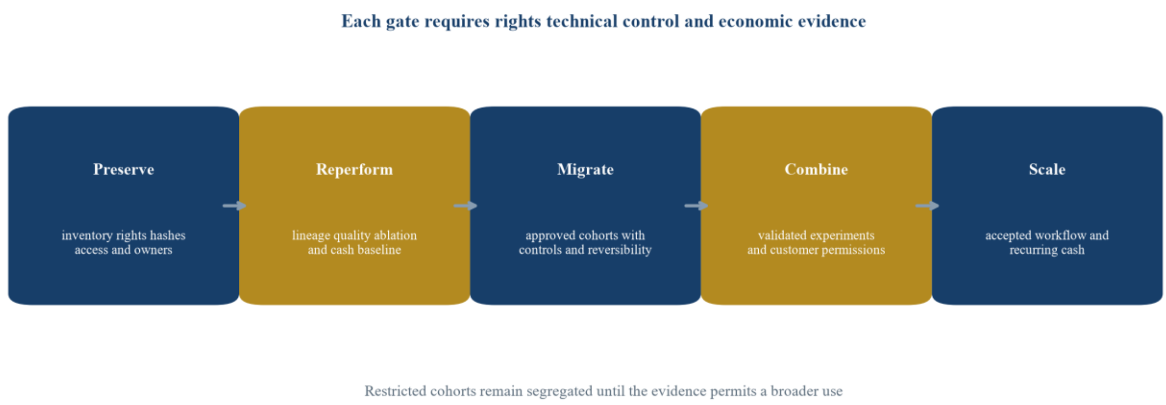
The first phase should freeze the inventory, record hashes, access controls, source terms, model versions, evaluation sets and responsible owners. Restricted cohorts should remain segregated. New access should follow least-privilege rules. The buyer should establish incident, deletion and customer-request procedures before expanding use.

The second phase should reproduce priority performance and economics. Teams should run agreed tests, confirm current refresh flows, reconcile costs, interview workflow owners and verify customer acceptance. This establishes a closing-date baseline against which integration effects can be measured.

The third phase should migrate selected data and pipelines through documented interfaces. Each migration should preserve lineage, purpose limits, retention rules and audit evidence. Combined-model experiments should use approved cohorts and locked evaluations. Results should be reviewed for performance, subgroup effects, privacy, security and commercial relevance.

The final phase should release value in increments. Product expansion, cross-customer learning, sublicensing or new geographies should proceed only when rights, technical performance, controls and customer acceptance support them. Integration incentives should reward recurring cash and control quality rather than raw data consolidation.

Figure 5 Evidence gated data integration sequence



Proposed sequence; timing and approvals depend on applicable law contracts technical risk and the transaction perimeter.

13 Establish board reporting and post-close impairment signals

The board should receive a compact dashboard that connects rights, quality, contribution, refresh and cash. A headline record count should be accompanied by eligible-record coverage, provenance confidence, label agreement, drift, cost per usable record, test performance, workflow adoption, customer concentration, unresolved claims and recurring cash attributable to the relevant products.

Thresholds should be defined before closing. A fall in eligible-record coverage can trigger remediation. Loss of a major source can prompt a valuation review. Rising refresh cost can reduce sustainable margin. A model upgrade that diminishes the dataset's incremental contribution can shorten economic life. Customer restrictions or regulatory findings can change permitted use.

Post-close reporting should preserve the distinction between observed facts and management estimates. Actual records, contracts, costs, test results and cash can be reported as evidence. Future adoption, synergies and option value remain assumptions until supported by accepted use and collected revenue. The board should see scenario ranges and the reasons for changes.

Accounting teams should monitor indicators relevant to recognised intangible assets and goodwill under the applicable reporting framework. Transaction valuation, purchase-price allocation and subsequent impairment are connected decisions with different purposes and standards. Professional advice is required for the actual accounts.

The audit trail should allow a later reviewer to understand why the buyer paid for the asset, which assumptions were critical, what evidence supported them and how outcomes compared with the acquisition case. This discipline improves integration decisions and future transactions even when initial estimates prove inaccurate.

Conclusion

Proprietary training data can create substantial value in an AI acquisition. The value arises from enforceable and renewable rights, usable quality, scarcity relative to practical alternatives, measurable model contribution, operational adoption and recurring cash. None of these features can be established by record count or historic expenditure alone.

A defensible valuation begins with a component inventory and rights chain. It reduces headline volume to usable data, tests uniqueness against substitutes, reproduces incremental contribution, models decay and refresh, reconstructs sustainable earnings and reconciles cost, income and market indications. It then converts remaining uncertainty into price, escrow, indemnity, earn-out, closing conditions and a gated integration plan.

The result is an evidence-based transaction decision. It allows a buyer to distinguish a durable data production system from a costly archive, a transferable advantage from a customer-specific permission and demonstrated cash from an untested option. The same framework gives sellers a practical route to readiness: document provenance, clarify rights, isolate contribution, fund refresh and connect technical performance to accepted customer outcomes.

Frequently asked questions

Is a large training dataset automatically valuable

No. Size can improve coverage and model performance, but value depends on permitted use, quality, uniqueness, representativeness, refreshability, incremental contribution and cash. Duplicate, stale, restricted or poorly labelled records can increase cost without increasing economic benefit. The buyer should reconcile gross volume to usable records and then test contribution.

Can historic collection cost be used as the valuation

Historic cost is evidence, not a complete conclusion. It can include failed work, inefficient processes and expenditure that a rational buyer would avoid. A replacement-cost analysis should estimate the current cost and time required to create equivalent utility with lawful rights and appropriate controls. Income and market evidence should also be considered where available.

How should the buyer separate data value from model value

The buyer can use ablation tests, alternative datasets, learning curves and with-and-without cash-flow scenarios. The analysis should hold model architecture, compute, tuning and evaluation conditions sufficiently constant to isolate the data effect. It should also apply contributory-asset charges for software, workforce, customer relationships, brand, compute and other assets.

What happens when source rights are incomplete

The affected cohort should be identified and segregated. The valuation can exclude it, probability weight its contribution, include remediation or relicensing cost, shorten economic life or place consideration into escrow or an earn-out. Transaction counsel should determine representations, indemnities, closing conditions and remedies.

How should refresh cost affect value

Refresh cost reduces sustainable earnings and influences economic life. The model should include continuing acquisition, annotation, adjudication, privacy, security, documentation, retraining and validation. A dataset with a renewable source and efficient feedback loop may retain value longer than a static historical corpus.

Is a price per record a reliable market method

It is rarely sufficient by itself. Records differ in rights, exclusivity, domain, quality, context, rare-event density, recency and permitted use. Comparable evidence should be adjusted for these characteristics and for whether the transaction includes refresh access, labels, software, models or services.

Can synthetic data replace proprietary training data

Synthetic data can support augmentation, testing, privacy controls or rare scenarios. Its value depends on how it is generated, validated and linked to real-world performance. It may reduce reliance on some source data while preserving dependence on real observations for calibration, validation and drift detection. The buyer should test the best realistic synthetic alternative rather than assume equivalence.

Which post-close metrics matter most

Useful metrics include eligible-record coverage, provenance confidence, label agreement, subgroup performance, drift, cost per usable record, refresh cycle time, unresolved rights, customer retention, accepted workflow use and recurring cash. Thresholds should be linked to remediation, capital release and valuation review.

References

[1] World Intellectual Property Organization. WIPO supports aligning valuation standards with the modern economy. 2026. https://www.wipo.int/en/web/ip-financing/w/news/2026/wipo-supports-aligning-valuation-standards-with-the-modern-economy

[2] World Intellectual Property Organization. Valuing intellectual property assets. https://www.wipo.int/en/web/business/ip-valuation

[3] World Intellectual Property Organization. Intellectual property valuation in technology transfer. https://www.wipo.int/en/web/technology-transfer/access-market

[4] World Intellectual Property Organization. Frequently asked questions on AI and IP policy. https://www.wipo.int/en/web/frontier-technologies/artificial-intelligence/faq

[5] IFRS Foundation. IAS 38 Intangible Assets. https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/

[6] IFRS Foundation. IFRS 3 Business Combinations. https://www.ifrs.org/issued-standards/list-of-standards/ifrs-3-business-combinations/

[7] IFRS Foundation. IFRS 13 Fair Value Measurement. https://www.ifrs.org/issued-standards/list-of-standards/ifrs-13-fair-value-measurement/

[8] IFRS Foundation. Module 18 Intangible Assets other than Goodwill. https://www.ifrs.org/content/dam/ifrs/supporting-implementation/smes/module-18.pdf

[9] European Union. Regulation EU 2016/679 General Data Protection Regulation. 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[10] European Union. Directive EU 2019/790 on copyright and related rights in the Digital Single Market. 2019. https://eur-lex.europa.eu/eli/dir/2019/790/oj

[11] European Union. Directive 96/9/EC on the legal protection of databases. 1996. https://eur-lex.europa.eu/eli/dir/1996/9/oj

[12] European Union. Directive EU 2016/943 on protection of undisclosed know-how and business information. 2016. https://eur-lex.europa.eu/eli/dir/2016/943/oj

[13] European Union. Regulation EU 2023/2854 on harmonised rules on fair access to and use of data. 2023. https://eur-lex.europa.eu/eli/reg/2023/2854/oj

[14] European Union. Regulation EU 2024/1689 laying down harmonised rules on artificial intelligence. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[15] United States Copyright Office. Copyright and Artificial Intelligence. https://www.copyright.gov/ai/

[16] United States Copyright Office. Copyright and Artificial Intelligence Part 3 Generative AI Training pre-publication version. 2025. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

[17] European Data Protection Board. Opinion 28/2024 on data protection aspects related to processing personal data in AI models. 2024. https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en

[18] European Data Protection Board. GDPR principles support responsible AI. 2024. https://www.edpb.europa.eu/news/edpb-opinion-on-ai-models-gdpr-principles-support-responsible-ai_en

[19] United Kingdom Information Commissioner's Office. Guidance on AI and data protection. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

[20] United Kingdom Information Commissioner's Office. Explaining decisions made with AI. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

[21] European Commission. Data Act explained. https://digital-strategy.ec.europa.eu/en/policies/data-act

[22] European Commission. Data Act model contractual terms and standard contractual clauses. https://digital-strategy.ec.europa.eu/en/policies/data-act-model-contractual-terms-and-standard-contractual-clauses

[23] European Commission. AI Act regulatory framework. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[24] European Commission. Common European data spaces. https://digital-strategy.ec.europa.eu/en/policies/data-spaces

[25] European Commission. Competition policy and data. https://competition-policy.ec.europa.eu/sectors/digital-media-and-internet/data_en

[26] European Commission. EU merger control. https://competition-policy.ec.europa.eu/mergers/overview_en

[27] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework 1.0. 2023. https://doi.org/10.6028/NIST.AI.100-1

[28] National Institute of Standards and Technology. Generative Artificial Intelligence Profile. 2024. https://doi.org/10.6028/NIST.AI.600-1

[29] National Institute of Standards and Technology. AI RMF Playbook. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

[30] National Institute of Standards and Technology. AI Resource Center. https://airc.nist.gov/

[31] International Organization for Standardization. ISO IEC 42001 artificial intelligence management systems. https://www.iso.org/standard/81230.html

[32] International Organization for Standardization. ISO IEC 23894 artificial intelligence risk management. https://www.iso.org/standard/77304.html

[33] International Organization for Standardization. ISO IEC 27001 information security management systems. https://www.iso.org/standard/27001

[34] International Organization for Standardization. ISO IEC 5259 data quality for analytics and machine learning. https://www.iso.org/standard/81088.html

[35] OECD. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[36] OECD. OECD AI Principles. https://oecd.ai/en/ai-principles

[37] European Union Agency for Cybersecurity. Data protection engineering. https://www.enisa.europa.eu/publications/data-protection-engineering

[38] National Institute of Standards and Technology. Cybersecurity Framework 2.0. 2024. https://doi.org/10.6028/NIST.CSWP.29

[39] National Institute of Standards and Technology. Privacy Framework. https://www.nist.gov/privacy-framework

[40] National Institute of Standards and Technology. Secure Software Development Framework. https://csrc.nist.gov/Projects/ssdf

[41] Data Management Association International. Data Management Body of Knowledge. https://www.dama.org/cpages/body-of-knowledge

[42] International Valuation Standards Council. International Valuation Standards. https://www.ivsc.org/standards/

[43] International Valuation Standards Council. Perspectives on intangible assets. https://www.ivsc.org/intangible-assets/

[44] Financial Accounting Standards Board. Business combinations topic 805. https://asc.fasb.org/topic&trid=2120890

[45] United States Securities and Exchange Commission. Regulation S-K. https://www.ecfr.gov/current/title-17/chapter-II/part-229

[46] Federal Trade Commission. Artificial intelligence and algorithmic tools. https://www.ftc.gov/business-guidance/technology/artificial-intelligence

[47] United States Department of Justice and Federal Trade Commission. Merger Guidelines. 2023. https://www.justice.gov/atr/2023-merger-guidelines

[48] European Commission. Guidelines on the definition of an artificial intelligence system. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application

[49] World Intellectual Property Organization. World Intangible Investment Highlights 2026. https://www.wipo.int/web-publications/world-intangible-investment-highlights-2026/en/key-trends-and-insights.html

[50] World Intellectual Property Organization. Technology Trends Artificial Intelligence. https://www.wipo.int/publications/en/details.jsp?id=4386

About the Author

Chennakeshav (CK) is a corporate finance and investment banking executive with 25+ years of global experience in deal origination, structuring and execution across M&A, growth capital and corporate strategy. He has led value-creation mandates for founders, corporates and funds — bridging the boardroom view to hands-on execution and close.

His career spans Morgan Stanley, HSBC, Lloyds Banking Group, EWEC, ADQ portfolio companies and Emirates Growth Fund, across TMT, real estate, fintech, deeptech, cleantech, infrastructure and energy. He has partnered with C-suite leaders, private equity and venture funds, sovereign wealth funds and family offices to finance complex fund raises and scale-up ventures, and has led M&A due diligence, post-merger integration and business-transformation initiatives to create value.

At Matchpoint Partners he is Managing Partner, leading the firm's corporate finance, M&A and capital-raising practice. He holds an MBA from London Business School, an engineering degree from VTU and a Master of Laws (LLM, in progress) from UCL London.

An active start-up mentor, CK mentors at Techstars, DIFC FinTech Hive, Startup Grind, Founder Institute and IN5, serves as Entrepreneur Mentor in Residence (EMiR) at London Business School, and judges the Entrepreneurship World Cup.

https://www.linkedin.com/in/ckadya/

https://www.matchpoint-partners.com/team/ck-adya.html

This paper is part of a continuing series on the structure of private and alternative markets. The views expressed are the author's own. The paper is for information only, describes market structure in general terms, and does not constitute investment, legal, tax or regulatory advice or a recommendation in respect of any security, vehicle or counterparty.