

Mechanisms for a Sustainable AI Data Economy

Background Note on AI Data Markets (v3)

September 15, 2026

Ilan Strauss, Anjali Shrivastava, and round table organizers¹

Note: This piece focuses largely on the data being consumed by AI models. For more on mechanisms, protocols, and business models for shared value creation and distribution connected to data & AI native artifacts please see: [here](#).

Table of Contents

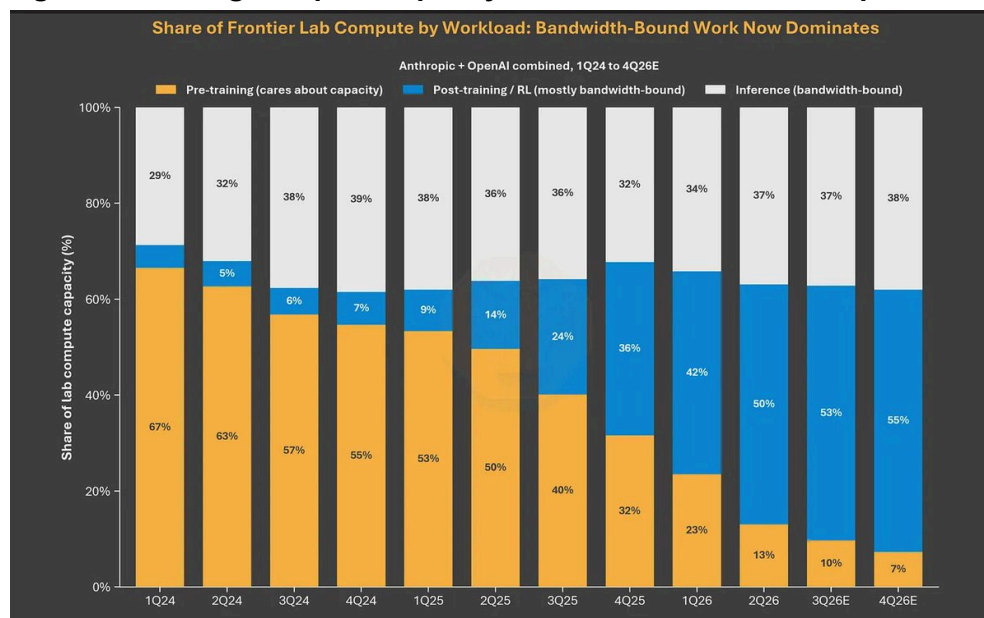
AI Needs Data.....	1
Data Types: What is actually being bought and enclosed?.....	1
The Challenge of Synthetic Data: Derived products.....	3
How is Data Being Procured or Enclosed?.....	5
Pricing: How much?.....	6
Who participates, who benefits?.....	7
From deals to markets.....	8
Institutions and Licensing terms.....	9
Access signals and enforcement.....	9
Market infrastructure and public rules.....	10
Valuation and Attribution Mechanisms.....	11
Counting rules.....	11
Contribution rules and Shapley values.....	11
Training-time attribution issues.....	12
Market gaps.....	12
Some Open Questions.....	13

¹ Initial draft by Anjali Shrivastava. Edited and extended by Ilan Strauss and the convening organizers. AI has been used in the writing and research process. All material has been read, edited, and checked by a human. Errors may remain.

AI Needs Data

Labs need data to train and evaluate increasingly capable frontier models. But the required inputs are becoming more varied as models are pushed into new tasks and domains. “Data” now includes expert demonstrations and recorded work sessions, real work outputs, many-step tasks, reinforcement-learning (RL) environments, proprietary data collections, private evaluations, and records of the steps a model takes when people use it to do difficult work.

Figure 1. Shifting compute capacity at Frontier Labs towards post-training and inference



Source: [SemiAnalysis](#).

Note: “end token demand has increased and post-training techniques became a major vector for model scaling.”

Data Types: What is actually being bought and enclosed?

“Data” is not a single economic good. It encompasses assets and production processes that differ in scarcity, cost, ownership, reproducibility, refresh rate and how they create value for AI systems.

- **Proprietary knowledge and context:** Books, archives, code, business records and other accumulated information that models can draw on. Value depends on things like scarcity, authority, recency and the availability of substitutes.

- **Interaction trajectories:** Records of *how tasks are actually performed*, including: prompts, searches, tool calls, edits, intermediate outputs, failures, corrections and eventual outcomes. Unlike a static document, a trajectory reveals the process by which knowledge is turned into action.
- **Human expertise and judgment:** Demonstrations, rankings, annotations, critiques and solutions supplied by people with relevant skills. The scarce input is often not the resulting dataset itself, but access to people capable of repeatedly producing high-quality judgments.
- **Synthetic and simulated experience:** Training material generated by models or physical simulators, rather than directly observed in the world. Real data or trajectories can serve as seeds from which much larger quantities of synthetic tasks, solutions and experiences are produced.
- **Evaluation and reward signals:** Benchmarks, graders, tests and other mechanisms for determining whether a model has succeeded. They provide the feedback necessary to improve model behaviour.
- **Data-generating capacity:** Continuing access to users, workers, firms, institutions or environments that produce new information over time. This can be more valuable than a fixed corpus because it creates a renewable stream of proprietary context, trajectories and feedback.

What's included in the contract – which bundle of rights? Rights can be separated, into among other things: permission to pre-train, post-train, evaluate, retrieve or ground content at inference time (that is, bring external sources into the answer while the model is generating it), summarize or display content, retain copies, create derivative datasets or products, resell or sublicense, use exclusively, and require attribution, usage reporting or audit. These rights are economically – and often legally – distinct even though a contract may bundle them together.

Box 1. The value of interaction traces and trajectories

An interaction trace is a record of the individual actions and events generated while a person or AI system performs a task. This includes prompts, clicks, tool calls, searches, edits, observations and intermediate outputs. A trajectory is the ordered sequence of those interactions from the beginning of a task through to an outcome, including failed attempts, corrections and branching steps. Together, they record the process that produced the final work product.

Vertical integration or training on its users' flow data is one means an AI company can acquire privileged access to these traces by embedding its systems directly into a workflow. Those real traces can then be used directly for training and evaluation, or as seeds for generating synthetic tasks and trajectories at much larger scale.

The distinction between a stock and a flow is important. A static archive of legal documents is a

finite asset whose relevance can decay (a stock). A law firm, workflow platform, or expert network can continuously produce new tasks, drafts, corrections and feedback (a flow). Control of the production process (through vertical integration and acquisitions) to attain user trajectories can therefore be the strategy of choice for AI companies looking to maximize the value of data for continuous AI model improvement.

Economically, this shifts the question from a one-off pre-training data license or payment towards ongoing access, revenue sharing, portability and competition over the continuous flow of data that can guide model improvement and specialization.

The Challenge of Synthetic Data: Derived products

Increasingly, training data is generated rather than collected. Models can produce new problems, conversations, code, demonstrations and interaction trajectories at enormous scale. Adaption's new [Invent a Dataset](#) product, for example, lets a developer specify the dataset they want and generates structured, training-ready data rather than starting from an existing corpus.

Synthetic data is particularly powerful when answers can be independently verified. Code can be executed and mathematical answers checked. In physics, researchers have generated question-answer pairs by randomly creating scenes inside physics engines and training models entirely on those simulated interactions. [One 2026 study](#) reported that this improved performance on real International Physics Olympiad problems by *5-10 percentage points*, despite using synthetic simulation data for training. Simulators are not perfect though as models can learn impossible things.

Synthetic data can be used to make real data more useful for training. Real-world data often contains valuable information in a form that is difficult for models to learn from directly. Synthetic generation can transform that information into clearer, more structured, and more training-friendly examples while preserving the underlying knowledge. Microsoft's Phi models provide a prominent example. For Phi-1, Microsoft combined 6 billion tokens of "textbook-quality" web data with 1 billion tokens of synthetically generated textbooks and exercises based on that material. Microsoft later described this approach as useful because synthetic data can present reasoning in a more gradual, structured form and can rewrite facts from sources such as web forums into formats closer to the conversational contexts in which a model will ultimately use them. More recent Phi work similarly starts with filtered real-world questions and then synthetically rewrites or transforms them to target particular reasoning skills.

Synthetic data faces a diversity constraint. [Andrej Karpathy argues](#) that model-generated datasets can suffer from "silent collapse": individual examples look plausible, but collectively cover only a narrow portion of the underlying distribution. Generating more examples from the same model and prompts therefore does not necessarily create more information. The challenge is to preserve enough entropy and diversity to explore genuinely different parts of the task distribution (rather than repeatedly sampling variations of what the model already knows).

Simulation plays a similar role in robotics, where agents can attempt the same task thousands of times without the cost of collecting equivalent real-world demonstrations. One limitation is that synthetic data inherits the capabilities and errors of the system that generates them: an imperfect simulator can teach physically impossible behavior, while model-generated data can reproduce or amplify gaps in its seed data.

Biodiversity policy and collective benefit-sharing. Genetic sequence information can be transformed into commercially valuable drugs, crops and biotechnology products, often making it difficult to connect a final product to a particular original genetic resource. *The Convention on Biological Diversity* has therefore created the [Cali Fund](#), under which large companies commercially benefiting from digital sequence information are expected to contribute based on revenue or profit rather than attempting to price every individual sequence used. This deliberately substitutes *collective benefit-sharing for perfect attribution*. To date it has been a failure, however, with little payments coming in.

Distillation and training on model outputs. Synthetic data has been used to quickly make a smaller model mimic a larger model by training on its output. This is particularly valuable in the post training stage where a smaller model can learn to copy how the model writes or thinks without training the model from scratch. Distillation becomes more powerful when the token probabilities of a model's output are exposed because it allows the student model to learn the full distribution of probabilities from the teacher model at once.

Box 2. Derived products under law

Synthetic data poses the question of if proprietary documents, human expertise or interaction trajectories are used to generate a much larger synthetic dataset – and that dataset is then used to train a valuable model – at what point, if ever, do claims associated with the original inputs end?

Legal systems have long had to decide what happens to ownership when one asset is transformed into another. Roman law's doctrine of [specificatio](#) famously addressed who owned a new object made from somebody else's materials. Justinianic law adopted a compromise based on reversibility: if the new thing could be reduced back to its original material, ownership generally remained with the owner of the material. Modern law however looks not just at physical transformation but economic contributions: ownership aside, whose contribution predominantly accounts for the final output's value?

Modern music copyright similarly allows new works to generate rights of their own while preserving rights in the underlying composition or recording; while U.S. copyright law defines derivative works as works in which preexisting works are recast, transformed, or adapted. But purely statistical influence without incorporation of protected expression is a murkier case.

Sources: Wikipedia, ChatGPT, and Barry Nicholas, *An Introduction to Roman Law* (Oxford: Clarendon Press, 1962), pp. 136-138.

How is Data Being Procured or Enclosed?

Labs are procuring and enclosing data inputs through several channels:

Expert-data marketplaces and reinforcement-learning (RL) environment suppliers: *Scale AI*, *Surge AI*, *Mercor*, and *Handshake* recruit and manage expert contributors who produce post-training data, evaluations and task environments. Here post-training means additional training applied after a base model has already been pretrained. *Mechanize* focuses more directly on realistic simulated work environments, tasks and graders that score model performance on long-horizon (many-step) software work. The boundary between a “data vendor,” labor marketplace, and environment builder is increasingly blurred.

One-off purchases of corporate archives: Google [agreed](#) in August 2026 to pay \$10 million for internal data from bankrupt Spirit Airlines, including employee emails, Teams messages, spreadsheets, calendars and operational data, for product development and AI training. *Mercor* separately submitted a \$7.5 million bid.

Data brokerage: [Handshake](#) now offers companies anything from \$100,000 to \$4 million to license internal company data, which it de-identifies by removing or masking identifying information and delivering to frontier labs for training and evaluation. Its examples include email archives, company drives, enterprise resource planning (ERP) systems and customer relationship management (CRM) data.

Vertical acquisition of context: A vertically integrated strategy is for the AI company to acquire or enter into organically the businesses that generate valuable domain-specific data in the course of their work. Anthropic’s Claude model providing a vertical add-on for finance or design is one such example. The *Context Acquisition Company* – a new holding company – explicitly aims to buy service firms partly for the proprietary data and expertise embedded in their operations, and use that context to build specialized models and agents.

Private evaluations: Many evaluation companies build proprietary benchmarks from non-public datasets in law, finance and other domains, creating another market for scarce information about what models can and cannot do. This is especially relevant amid calls for third party evaluation, and some have [speculated](#) that these independent evaluators are another channel for labs to gather high-quality data.

Licensed vertical knowledge and inference-time access: A notable example is that of *OpenEvidence*, a clinical AI platform that has signed a multi-year licensing agreement with the *JAMA Network*, giving it access to full-text and multimedia content from JAMA and its specialty journals to inform answers generated by the platform. *Harvey* – a widely used AI for lawyers – integrates licensed authoritative sources from LexisNexis, Wolters Kluwer and Otto Schmidt, and litigation evidence through Everlaw via MCP (forthcoming). Rather than acquiring a static training corpus, [Harvey](#) buys access to continuously updated legal sources while the model is

answering a query. It also states that customer data and content [are not used to train](#) shared AI models.

Usage- and contribution-linked licensing: [Parallel AI](#)'s Index product used for AI search and research during *inference* estimates each website source's marginal contribution to an agent's work using a Shapley-value approach (explained below), and pays more for harder-to-replace content. The aim is to price source-specific "alpha": uniqueness rather than volume or citation frequency. Google's 2026 "[AI contribution](#)" pilot also makes recurring payments when Google judges publisher content to have contributed to Gemini, AI Overviews or AI Mode responses; Google does not disclose the valuation formula.

Licensing and payment infrastructure: [RSL](#) 1.0, a machine-readable licensing standard, lets publishers express machine-readable terms such as attribution and usage-based payment for each AI answer that uses the source; [Cloudflare's Pay Per Crawl](#) charges at the access layer; [TollBit](#) offers summarization and full-display licenses priced per thousand pages. These tools aim to reduce contracting and payment frictions (so-called "transaction costs").

Pricing: How much?

Most transaction prices for data are private. The public record remains heavily seller-side because even listed AI developers rarely disclose data acquisition as a separate expense, rather listing it under R&D, third-party services, or other line items.

The figures below are disclosed or credibly reported public reference points, not a market-size estimate. They are also not directly comparable: some are licensing fees, some are revenue pools that are shared, some are labor-marketplace revenue, and one is a legal settlement.

Table 1. Value of Existing Data Agreements / Purchases

Channel	Seller / program	Reported price or scale	Pricing unit
Expert data	Mercor	More than \$2bn gross annualized revenue (June 2026); reported 60–70% paid to contractors	Expert hours
Expert data / supplier scale	Scale AI	Meta paid \$14.3bn for a 49% non-voting stake (June 2025)	Equity, not a price for data
Synthetic Data	Gretel → Nvidia	Reported acquisition price exceeds Gretel's last known valuation of \$320m; terms undisclosed (March 2025). Gretel had raised \$67.7m pre-deal.	Equity / M&A, not a price for data

User content	Reddit	\$203m aggregate value across contracts signed in Jan. 2024, over two- to three-year terms	API access + quarterly data transfers
News	News Corp → OpenAI	Reportedly could exceed \$250m over five years, including cash and OpenAI credits potentially	Archive/current content for training and answers
Academic publishing	Wiley	\$40m AI-licensing revenue in FY2025; \$23m in FY2024	Per-project licensing
Academic publishing	Informa / Taylor & Francis	>\$75m total AI-partnership revenue in 2024 across two agreements; Microsoft agreement reportedly >\$10m in first year plus recurring fees.	Access fee + recurring payments
Trade books	HarperCollins (buyer reported as Microsoft)	\$5,000 per title for three years; opt-in; payment split equally between author and publisher	Per title
Legal benchmark	Bartz v. Anthropic settlement	\$1.5bn for 482,460 works, about \$3,000 per work before fees	Litigation settlement, not a negotiated license price
Grounding / inference	Perplexity Comet Plus	80% of Comet Plus subscription revenue allocated to participating publishers; \$42.5m initial pool	Revenue share using traffic, citation and agent-use signals
Grounding / inference	Google AI contribution pilot	Undisclosed; participating publishers receive a monthly payment	Payment based on Google-judged contribution
Crawling	Cloudflare Pay Per Crawl	Publisher sets a price for an AI crawler request	Per crawl

Source: Various see hyperlinks.

The disclosed figures show that commissioned expert work is already a billion-dollar channel, while disclosed publisher licensing is economically material but much smaller in the visible record.

Who participates, who benefits?

- **The long tail on the web often blocks AI rather than sells, with the rest participating via social media platforms.** A [2024 study](#) of C4, a large web corpus used in language-model research, found that by April 2024 robots.txt restrictions applied to

about 5% of all text tokens (small pieces of text used by models) and more than 28% of its most actively maintained, important sources; terms of service restricted about 45% of the corpus. In practical terms, many smaller sites have a way to refuse access but no functioning market in which to state a price. Smaller content creators can participate via social media platforms, which offer an alternative to the open web business model invented by Google Search and its pay-per-click advertising.

- **The licensor and the contributor are often different parties.** A publisher, platform or employer may sign the contract, while authors, posters, employees or customers receive payment only if their own contract with the licensor passes value through. *HarperCollins'* reported opt-in author split is one model; other deals have provided no individual opt-out or stated creator share.
- **Expert marketplaces broaden supply, but then mediate the relationship.** *Mercor* reports more than 30,000 weekly active contractors and more than \$2m a day in contractor payments. Experts can participate without negotiating with a frontier lab, but the marketplace controls recruitment, quality assurance, contract terms and visibility into downstream use.
- **Buyers remain concentrated.** Frontier labs account for much of reported demand, but application companies are becoming buyers too. *Harvey* licenses authoritative legal content and uses metered web access; companies such as *Notion* and *Opendoor* have also appeared as customers of new inference-time data services.

From deals to markets

Enterprise customers are increasingly structuring their AI workflows so that proprietary information does not become training data fodder for another company. Palantir, Nvidia and others have [reportedly](#) restricted some use of frontier-lab AI services because of concerns about data retention and leakage of proprietary information.

Deals in these overlapping “AI data markets” are often bilateral and relationship-driven. Prices are rarely public, and the methods used to set those prices are even less visible.

One question for this roundtable is what would have to change for this collection of bespoke deals to become more transparent, liquid public markets with broad participation and wider distribution of value.

Note that “liquid” here does not mean a commodity exchange with one spot price for data. The goods are too heterogeneous and buyer-specific for that. Here, greater liquidity means better discovery, comparable descriptions of what is being sold, lower verification and contracting costs, credible quality signals, more reliable price discovery, and mechanisms that let smaller suppliers participate without negotiating every transaction from scratch.

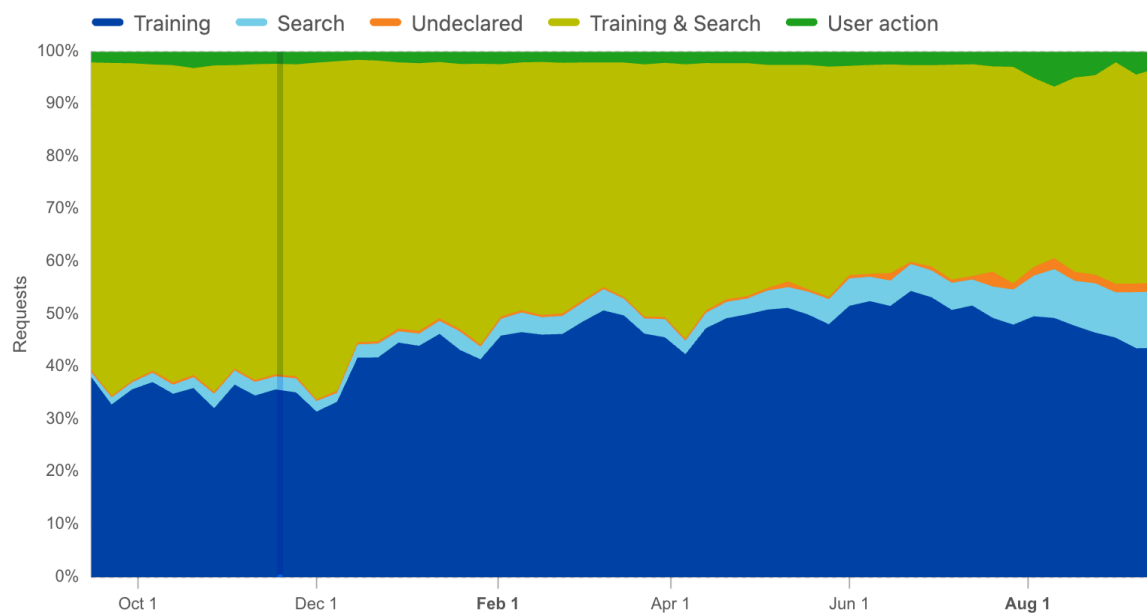
Institutions and Licensing terms

Institutions, as devices for structuring and enforcing collective behaviour, are relevant and often necessary devices for advancing new mechanisms (rules) that align incentives around new value creation propositions. Pieces of market infrastructure are emerging, but no single institution currently handles access, rights, valuation, audit, and payment to underlying contributors.

Access signals and enforcement

- **Robots.txt and crawler controls.** A robots.txt file is a public instruction file that tells automated web crawlers which parts of a site they may access. AI companies increasingly now use named crawler tokens, such as Google-Extended, so a publisher can distinguish some AI uses from ordinary web search. The limitation is that one token may not cover every product or downstream use.
- **Machine-readable licensing terms.** RSL 1.0 extends the access-control idea by allowing publishers to state licensing conditions in a machine-readable form, including attribution and pay-per-crawl or inference. The RSL Collective can act as a licensing agent for participating rightsholders.
- **Network-level payment and enforcement.** Cloudflare can block known AI crawlers for sites using its network. Its Pay Per Crawl mechanism can answer an unpaid crawler request with HTTP status 402 (“Payment Required”) and a price, turning an access request into a potential transaction rather than a simple allow/block decision.

Figure 2. AI bot crawl purpose by HTTP traffic by type (worldwide, September 2025-September 2026)



Source: [Cloudflare Radar](#), September 15, 2026 (12-month view). The categories distinguish bots used for model training, search, both, and user-triggered actions.

Market infrastructure and public rules

Inference-time marketplaces. Microsoft's [Publisher Content Marketplace](#) lets publishers set terms for content used to ground AI answers and provides usage reporting. "Grounding" means retrieving external content at the time a model answers a query, rather than relying only on information learned during training. *Parallel's Index*, *Human Native*, and *Google's pilot* are other attempts to intermediate this market.

In most inference-time programs, the same buyer or intermediary that routes the content also measures usage or contribution. That lowers transaction costs but creates a governance issue: sellers depend on a counterparty's measurement system to determine what they are owed.

Disclosure rules. Under [Article 53\(1\)\(d\)](#) of the EU AI Act, providers of general-purpose AI models must publish a summary of the content used for training using a European Commission template. This does not reveal transaction-level prices or attribution, but it creates a baseline disclosure obligation about training inputs.

Courts may initially become de facto price-setters. When licensing institutions fail, litigation can establish large monetary benchmarks. The [Bartz v. Anthropic](#) settlement is not a market price for a license, but it provides a public reference point for the cost of disputed acquisition practices.

Mechanisms

Some principles:

Turn use into a measurable event. YouTube's Content ID detects when copyrighted music is reused and records that use against a known rightsholder, who can then monetize, track, or block it. This converts difficult-to-observe reuse into a recordable event that can support enforcement and payment.

Reward quality, not just willingness to pay. Google Ads does not rank ads purely by bid amount. Ad Rank also incorporates expected click-through rate and relevance (including landing-page quality and other metrics). Advertisers, therefore, have an incentive to improve what users actually see rather than simply bid more. Conversion tracking also makes downstream outcomes more measurable.

Screen for quality before participation. Expert-data platforms already score and screen for quality. These mechanisms help match difficult work to contributors capable of producing it and prevent low-quality submissions from crowding out higher-quality contributions. *Mercor*, for example, requires applicants to complete assessments before matching them to projects, while *Scale* applies task-, dataset- and contributor-level quality checks.

Reward difficult-to-replace contributions. *Parallel's* Shapley-style approach tries to measure a source's (average) marginal contribution to an AI agent's output – its uniqueness or “[alpha](#),” as we called it – rather than simply counting citations or pages. In principle, this reduces incentives to flood the market with redundant material and rewards genuinely distinctive information.

By way of precedent: *YouTube Content ID* provides infrastructure for identifying uses of copyrighted music and administering the rights and payments attached to them. Music rightsholders supply reference recordings and metadata to YouTube, which creates digital fingerprints. YouTube then scans uploaded videos for matches. The system can represent different rights in a song, including the sound recording and shares in the underlying musical composition, and link matched uses to the relevant rights assets. Rightsholders can then choose to track, block or monetize those uses. Where they do choose to monetize, advertising revenue can flow to the relevant rightsholders and, in some arrangements, be shared with the uploader. Content ID links a detected use to a rights registry and, where monetized, to payment.

YouTube Content ID also provides infrastructure for tracking nested and fractional claims over a jointly produced output by linking a matched recording to the underlying composition rights and the ownership shares of multiple rightsholders

Valuation and Attribution Mechanisms

Pricing mechanisms divide roughly into two families. Counting rules pay for an observable event such as a crawl, citation, access request or expert hour. Contribution rules try to estimate how much a source changed the quality or economic value of an output. Counting is easier to verify; contribution is closer to the economic concept of [marginal value](#) but harder to measure.

Counting rules

Per-crawl, per-access, and revenue-share systems are comparatively cheap to administer because the event can be logged. Their weakness is that they may end up paying two sources the same amount even when one is redundant while the other contains valuable information that cannot be obtained elsewhere. Citation counts have the same limitation: a source can be economically valuable without being visibly cited, and a frequently cited source need not be causally important to the answer.

Contribution rules and Shapley values

A [Shapley value](#) is a standard way to divide the value created jointly by several contributors. Let N be the set of available sources, n the number of sources, and $v(S)$ the value produced when the system can use only the subset S . The Shapley value of source i is its marginal contribution averaged over every possible ordering in which sources could be added:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

The sum runs over every subset S of the other sources. The term $v(S \cup \{i\}) - v(S)$ asks the counterfactual question: *how much better is the result when source i is available?* The factorial

weight gives each possible position in the ordering equal importance. Exact calculation is expensive because it requires evaluating an enormous number of source combinations; with n sources, there are 2^{n-1} relevant source subsets for each source.

Parallel's Index applies this similar logic at inference time. It estimates how an agent's completed work (search output or deep research) changes when a URL or domain is removed as an input from its consumption, then uses learned approximations rather than calculating every coalition exactly. Its economic aim is to pay more when a source is unique (i.e. difficult to substitute) and used in a high-value task. This is a more demanding standard than paying per citation: the question is whether the source changed the result, not merely whether it appeared in the retrieval set. Similar ideas were advanced by a working group at Bellagio on Missing Mechanisms for the Human+AI Economy, inspired by a presentation from Alejandro Perez on the concept of "[alpha](#)".

Training-time attribution issues

At a model's training time, "influence functions" and related methods try to estimate how much particular training examples affected a model's behavior. In principle, that could support attribution after training. Some [useful proposals](#) in this area exist. In practice, the computation is difficult at frontier-model scale..

Quality is difficult to inspect before purchase. Buyers do not know how much a dataset will improve a model on X capability until it's been trained on. It is also hard to assess if an environment is robust to reward hacking, or if the underlying tasks are representative of real-world work.

Any attempt to value data before sale also runs into a classic information problem associated with Kenneth Arrow: the buyer often cannot know the value of information without seeing it, but once the buyer has seen it, the seller has already revealed part of what was being sold. Inference-time payment partly avoids this by pricing after use. It does not eliminate the trust problem though; moving it to the party that measures use and contribution. [Ablation](#) currently tries to measure the impact or uplift of a given data source to a model but is expensive to run.

Market gaps

1. *New business models are often absent and backward-looking.* These tend to focus on valuing existing outputs that have been consumed by humans on the web over the past two decades, rather than on outputs that must now be revalued and consumed primarily by machines; or on the new outputs that are being created using AI and are an entirely new class of market artifact.
2. *Multiple anchors for price.* Contracts are often denominated in production cost (i.e. hours), historical relationship between vendor and buyer, estimated improvement on X benchmark, scarcity of data, among other factors.

3. *There is no accepted basis for valuation.* Production cost, scarcity, benchmark improvement, access frequency, downstream revenue and marginal contribution can all produce different prices for the same source. Shapley-style methods are attractive because they target marginal value, but they require counterfactual measurement and a trusted estimator. A market also needs to decide whether the metric is merely informative or actually determines payment.
4. *Institutions are missing.* In practice, most value allocation is done through negotiation, i.e. through an institutional bargaining process on the basis of structured power relations, rather than solely through any measurable contribution to the socially produced surplus (even though this may be the limiting factor). Institutions also help unbundle and allocate rights, such as for compulsory licensing agreements, which has been central to dispersed markets arising where value can be broadly created and distributed.
5. *AI data's market structures risk being shaped by exclusivity and vertical integration.* Exclusive contracts and vertical ownership that continuously generate valuable data increase the incentives for AI companies to invest in specialized inputs. But this can be exclusionary and foreclose certain markets to potential competitors as important context becomes unavailable to rivals. It also reduces the scope for public valuation markets to arise and removes the possibility of non-rivalrous data sharing through open protocols.

Some Open Questions

1. What market structures, institutions, and mechanisms are needed when the valuable asset is a continuing flow of task traces and corrections generated through model use, rather than a static dataset sold once?
2. What usage report would allow a data seller to audit why they are receiving revenue and how their data was used during inference? And who should be trusted to compute it: the buyer, the marketplace, an independent auditor, or a shared open protocol?
3. How can we advance price discovery of data? What role should collective licensing play?
4. What role do independent evaluators play and what role should they play?