A Sustainable AI Economy Needs Data Deals That Work for Generators

Ruoxi Jia[*] Virginia Tech
Luis Oala[*] Brickroad
Wenjie Xiong Virginia Tech
Suqin Ge Virginia Tech
Jiachen T. Wang Princeton University
Feiyang Kang Virginia Tech
Dawn Song UC Berkeley

Abstract

We argue that the machine learning value chain is structurally unsustainable due to an economic data processing inequality: each state in the data cycle from inputs to model weights to synthetic outputs refines technical signal but strips economic equity from data generators. We show, by analyzing seventy-three public data deals, that the majority of value accrues to aggregators, with documented creator royalties rounding to zero and widespread opacity of deal terms. This is not just an economic welfare concern: as data and its derivatives become economic assets, the feedback loop that sustains current learning algorithms is at risk. We identify three structural faults—missing provenance, asymmetric bargaining power, and non-dynamic pricing—as the operational machinery of this inequality. In our analysis, we trace these problems along the machine learning value chain and propose an Equitable Data-Value Exchange (EDVEX) Framework to enable a minimal market that benefits all participants. Finally, we outline research directions where our community can make concrete contributions to data deals and contextualize our position with related and orthogonal viewpoints.

1 Introduction

Machine-learning at its core is a data processing chain: data shifts states from inputs to pre-train weights to synthetic outputs. The ascending adoption of AI products has commodified this value chain to a new level and triggered a land-rush for data (Figure 1). The market around this value chain is exploding. Model monetizers that have managed to transform data into a mercantile product see increasing sales. OpenAI alone reported $>$\$3.5 bn revenue in 2024 [1]. However, the distribution of that value is often lopsided. Papers, songs, or lines of code can be scraped, even pushing the envelope of legality [2], and distilled into revenue while data generators often receive little attribution. Data aggregators on the other hand, entities that amass large collections of data from generators, often benefit from the generous terms of service from pre-GPT times. Reddit, for example, banked \$203 million in data licenses until early 2024, yet channeled \$0 to the volunteers who wrote the content [3]. Data and its derivatives have become valuable commodities traded among a small cadre of firms, but the pipeline that transports value from data generators to model monetizers so far appears to be an extraction engine (Figure 2). In our view, the extraction is driven by three mutually reinforcing faults:

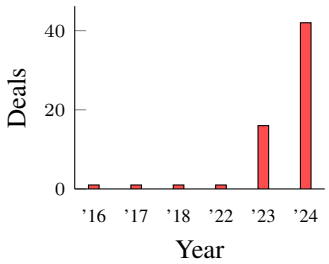

Figure 1: Recent data deals across past calendar years. See Appendix A for full list.

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

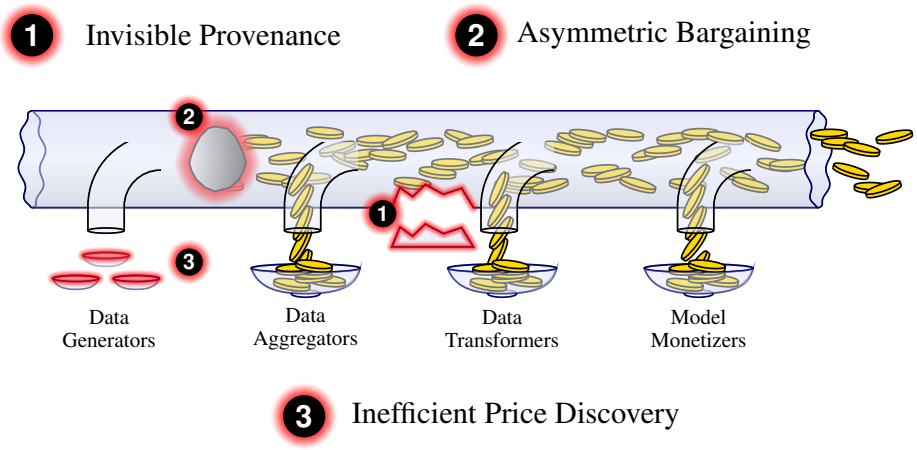

Figure 2: A pipeline symbolizing a piece of the data value chain in machine learning and the structural defects underlying the economic data processing inequality. 1) Data aggregators often strip provenance information of data generators when selling data to companies that transform data into model weights and monetizable products. 2) Model monetizers, which often also transform the data to model weights, enjoy a bargaining advantage as they control much of the current revenue generation. 3) Due to their heterogeneity, data generators in particular are not well equipped to participate in the price discovery of their own data.

invisible provenance, whereby pieces of data lose lineage and downstream users cannot audit or route royalties; **asymmetric bargaining power**, in which blanket licences and unilateral API terms let aggregators dictate prices to fragmented contributors; and **inefficient price discovery**, ignoring the dynamic, combinatorial value data accrues in different task contexts.

Taken together, our position is that the current ML value chain is unsustainable because of an economic data processing inequality that systematically transfers value away from data generators. A sustainable, efficient AI economy needs data deals that work for all market participants including generators.

If data generators are excluded from the value chain, the supply of high-quality, diverse data will shrink and prices will be set in opaque, concentrated markets. That dynamic stands to harm our own community: researchers, startups, and large labs alike risk facing fewer, less representative datasets—the very fuel current learning algorithms require. Conversely, we hypothesize that open, shared infrastructure facilitating data exchange between all market participants can help to stimluate data flow in an AI economy.

Contributions. **(1)** We compile and analyze a collection of 73 publicly disclosed data deals (Section 2). We trace how missing lineage, weak bargaining, and one-shot pricing form a feedback cycle that concentrates capital and cuts out data generators from the value chain. **(2)** In Section 3, we sketch an Equitable Data-Value Exchange Framework that wires task-data matching, dynamic data pricing, and auditable provenance into an efficient marketplace that benefits all participants. **(3)** We surface open problems—from scalable provenance tooling to incentive-compatible marketplaces—that our community can contribute to. **(4)** We weigh our observations and proposal against related (Section 4) and opposing view points (Section 5).

2 Economic Data Processing Inequality

Machine learning at its core is a data processing chain. While technical signal is carefully refined along this chain, economic equity is routinely removed from the original data generators—a circumstance we call *economic data processing inequality*. In our analysis we identify three structural mechanisms convolving into this processing inequality: *invisible provenance*, *asymmetric bargaining power*, and *inefficient price discovery*. These are not just concerns of economic welfare. As foundation models become economic actors or agents in their own right and data becomes a primary asset in that system of value creation, the link between data generation and weight harvesting that sustains today's learning algorithm is under strain. Capital concentration and market inefficiency compound with each generation of data derivative, threatening to disenfranchise data generators. Data aggregators enjoy

particularly powerful bargaining position at this time. This also raises the question whether data transformers, the actors refining data into model weights or labels, and model monetizers are getting the best deal compared to an open market. Figure 3 puts the scale and distribution of value in crass contrast: of the $677.3m in reported revenue, creator royalties round to almost zero. Symptomatically, 57 of the 73 found deals do not disclose any revenue publicly. The problem of dark figures appears widespread on both the number of deals transacted and their revenue volumes. A healthier market must work for model monetizers, data aggregators, and data generators alike.

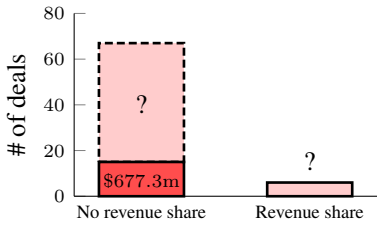

Figure 3: Counts of data deals with and wihtout revenue share information. Left bar (No revenue share): solid segment (15 deals) corresponds to deals where public sources quoted a revenue volume. If ranges are given we conservatively take the floor of that range. Total sum of disclosed revenue volume is $677.3m. The dashed segment (52 deals) indicates additional deals without revenue share and shows the problem of dark figures in this space. Right bar (Revenue share): 6 deals feature some revenue sharing information to generators, only one has public information on revenue ($2.5k).

Invisible Provenance. Once data is copied beyond its point of creation, contextual metadata—license, collection method, consent—often gets lost [4]. Provenance failures are by no means confined to academic benchmarks: they surface in healthcare ([5]), social media ([6]), or open-source code ([7]). As early as 2017, we have cases like the DeepMind–Royal Free breach where the UK Information Commissioner ruled that data generators "were not adequately informed that their data would be used" in deep learning products [8]. These types of disconnects between data's underlying license and its actual appeat to be a structural problem across the board and concerns many community datasets at large. An analysis on more than 1800 datasets has shown that more than 70% have license omissions and datasets with licenses have error rates of over 50% [4]. This causes provenance loss for data generators and legal uncertainty for data transformers and model monetizers. And the implications of this license uncertainty affect the value chain not only during training but also during inference on model weights. Across modalities, models can be shown to regurgitate images [9] and text [2] including content the model provider may not have the license for. In publicly disclosed license deals, the provenance disconnect between data and value generation already shows: a negligable fraction make provisions for revenue sharing with data generators (Appendix A and Table 2). Provenance gaps cascade through the value chain. The problem compounds once models are fine-tuned on material they themselves have produced [10, 11]. Already models train on a mix of original human content and synthetic content, which creates increasingly complex lineage graphs. Without robust, practical provenance a downstream user cannot audit permissions, enforce attribution, or route royalties, which in turn can chill secondary innovation and disrupt the feedback loop that rewards originators.

Asymmetric Bargaining Power. Even when provenance is intact, individual data generators are typically left out of deals between aggregators and model monetizers. Community platforms such as Reddit ($60 million per annum from Google [12]), Stack Overflow (bundled into Gemini [13]), and stock libraries like Shutterstock (multiple eight-figure licences in 2023–2024) all negotiate encompassing deals while offering creators click-wrap terms that grant sweeping reuse rights. Despite monitoring agencies such as the US Federal Trade Commission warning companies that such practices may be deemed deceptive [14], many platforms that host user-generated content (UGC) such as Google, Adobe, Snap, X or Meta have been modifying their terms of service regardless [15]. This power asymmetry also comes to bear in publicly disclosed data deals. More than 40% of known transactions were conducted by OpenAI/Microsoft, Google, or Perplexity on the buying side (Table 1), while aggregators bundle the receipts. These imbalances can snowball. Low marginal inference cost and winner-takes-most network effects channel control surplus to a handful

	Category	Deals
Top Types	News	26
	Images	16
	Academic	15
	UGC	14
Top Buyers	OpenAI	24
	Undisclosed	8
	Google	6
	Perplexity	3
Payment	Amount disclosed	16
	Undisclosed	57
	Recurring	3
	Generator split	6
	Litigation	4

Table 1: Aggregated snapshot on types, buyers and payment of 73 publicly disclosed data deals from Table 2.

of aggregators and monetizers. Maybe unintuitively this asymmetry also has the potential to harm the model monetizers who buy the data, because they negotiate with aggregator platforms rather than the generators directly in an open market. Rosen's "superstar" economics [16] suggests—and recent market caps of large AI companies confirm—that the top few firms stand to capture a lion's share of AI rents. Of the 73 transactions in Appendix A only six mention any revenue-share with contributors. In contrast, several deals, marked "L", are under litigation, signalling that they were hedged under the prospect of court action rather than negotiated at arm's length.

Inefficient Price Discovery. Where money does change hands it is almost always a lump-sum buy-out. News media licenses are illustrative: Associated Press agreed a two-year, flat-fee deal (amount undisclosed) [17]; News Corp settled for roughly $250 million across five years [18]; Axel Springer and Dotdash Meredith followed the same template. To the extent of public disclosure, these contracts do not appear to include royalty escalators for generators tied to usage, retraining, or downstream revenue. All of this plays out against a shifting legal backdrop: ongoing cases such as *Getty Images v. Stability AI*, *NYT v. OpenAI*, and *News/Media Alliance v. Cohere* hint that copyright doctrine, usage rights, and privacy statutes have yet to converge on generative training. Until clarity emerges, originators must choose between costly litigation and acquiescing to not participate in the market. In this sense, static payments are reinforced by the provenance gap. Once an originator has been taken out of the market equation, they have no financial stake anymore to participate in data quality, maintainenance of consent records, or police misuse, while the buyer internalizes much of the upside of any future innovation.

These three faults interact multiplicatively: missing provenance, undercuts bargaining, weak bargaining yields one-shot buy-outs, and buy-outs remove any incentive to invest in provenance. Breaking the cycle and creating an open market that maximizes overall welfare therefore requires technical and institutional interventions that address *all three dimensions at once*.

3 Towards an Equitable Data-Value Exchange Framework

The empirical cracks identified in Section 2 prompt the design of a data pipeline ensuring bargaining symmetry, provenance, and efficient pricing. This section outlines the Equitable Data-Value Exchange Framework (EDVEX), a minimal blueprint designed to empower diverse data contributors—especially smaller ones—and align market interests (Figure 5). We detail its potential layers and highlight key open problems, inviting discussion and future research.

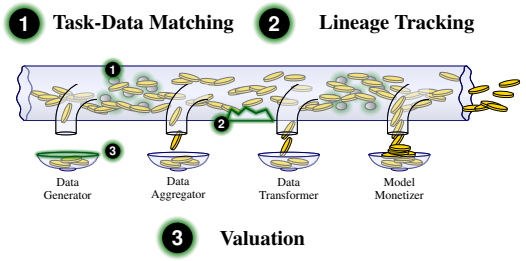

Figure 4: EDVEX patches for a sustainable and efficient machine learning economy.

3.1 Technical Primitives

Task-Data Matching. A foundational layer in an EDVEX would aim to optimally match data sources to specific machine learning tasks so as to maximize defined objectives for the model's effective task fulfillment [19]. In a vast and growing data landscape, especially one encouraging diverse contributors, automated discovery tooling becomes increasingly helpful. Simultaneously, data sellers—particularly smaller ones—often lack the resources or foresight to identify every context where their data might be valuable. This layer is foundational because it tackles the inefficiencies and information asymmetries that currently hinder effective data utilization. It ensures that data can be surfaced based on its utility for a task, rather than solely on the marketing capabilities or existing relationships of the data provider, thereby fostering a more level playing field and broader participation.

One might envision this as a sophisticated recommendation system for data, navigating the vast universe of potential sources to identify *combinations of sources* predicted to yield the highest improvement for a given target task [20, 21]. Traditional data discovery, often reliant on keyword searches, struggles to predict a dataset's true utility for downstream ML tasks [22]. Drawing inspiration from empirical scaling laws [23, 24, 25], which demonstrate that performance gains on representative data subsamples and small model sizes can help predict gains from the full corpus and larger model size, we could leverage a *sandbox-first evaluation protocol*. Within this sandbox,

candidate datasets (or their representative subsamples) are subjected to small-scale experiments (e.g., using small-scale proxy models) and the resulting performance changes are extrapolated to estimate their marginal utility for the target model at the desired data scale [20, 26, 27, 28]. Datasets are then ranked by this task-specific utility estimate, with the highest-scoring individual sources or bundles recommended to the developer. The sandbox can be designed to include additional features such as optimizing the mixture of high-ranked datasets, the implementation of which could build upon a growing body of research employing small-scale experiments to inform training data selection [29, 30, 31, 32]. The layer could also include data wrangling operations [33] (e.g., cleaning, transformation, and format standardization) before performing utility estimation.

Overall, by evaluating and potentially combining disparate data sources based on their collective utility for a task, this layer could facilitate the organic formation of *dynamic, task-optimized data unions*. These adaptive groupings, formed based on specific task requirements rather than pre-defined domains as seen in traditional data collectives [34], could empower contributors by creating value more efficiently, which in turn increases their bargaining power against large aggregators.

Open Problems for Task–Data Matching

Data profiling under constraints. How can we design a profile a data source in ways that capture its potential utility for specific AI tasks—facilitating better data discovery and matching—while preserving data contributor privacy and ensuring that the profile itself does not diminish the incentive for data acquisition by prematurely disclosing excessive value [35, 19]?
Task profiling for effective matching. How can AI task descriptions effectively articulate model-specific requirements—such as existing data summary, intended model architecture, whether training is from scratch or based on a pre-trained model—to guide the contribution of high-value, relevant data that demonstrably improves downstream model performance [36]?
Scalability of the sandbox protocol. How can the sandbox evaluation protocol (subsampling, lightweight model runs, utility extrapolation) be implemented to scale efficiently to potentially millions of datasets and thousands of tasks without incurring prohibitive compute costs or latency?
Generalization of utility estimation. Current scaling laws have mainly focused on certain data modalities, model architectures, and AI tasks. How well do utility estimates derived from sandbox evaluations generalize across different data modalities (tabular, time-series, graph), model architectures, and complex AI tasks (e.g., reinforcement learning)?
Feedback loops and adaptive data discovery. How can the discovery system incorporate feedback from actual downstream model performance (after full data acquisition and use) to continuously refine its utility estimation techniques for new tasks [37, 38, 39]?

Lineage Tracking and Auditable Provenance. To pay data generators according to their contributions, we must identify who provided data and how data is used. Currently, for AI training, such a lineage tracking or provenance mechanism is not widely adopted. Typically, a dataset comes with a license stating the restriction of using the data; however, the license does not help with tracking how the data is actually used. Existing lineage tracking [40, 41] requires manual effort to add the lineage metadata.

We need a framework to properly log the data source and the usage of data, so that the information can later be used for data valuation. Each asset (including dataset, trained models, and intermediate values) should have lineage metadata indicating what and how the source data impacts the asset. In the case where the dataset comprises data from different data creators, the metadata should faithfully record all the data sources [42], including small contributions. One challenge here is to have an encoding scheme that efficiently represents combinations of data sources from potentially millions of data creators. The second challenge is that in the AI pipeline the resulting model is influenced by how data is used in the workflow, e.g., data filtering decisions, feature engineering choices, training configurations (e.g., hyperparameters), architectural selections, and training randomness, making lineage tracking far more complex than traditional data processing pipelines. For example, different filters (or data curation mechanisms in general) lead to different data selection choices. Meanwhile, data practitioners might not want to reveal all the details of the workflow. Hence, the framework should consider what information to include in the metadata to make lineage tracking accurate enough without significant memory and execution time overhead.

Such a lineage tracking framework should also be designed to be easily deployed by practitioners. Practitioners typically use existing frameworks or APIs, such as PyTorch, for data processing and AI

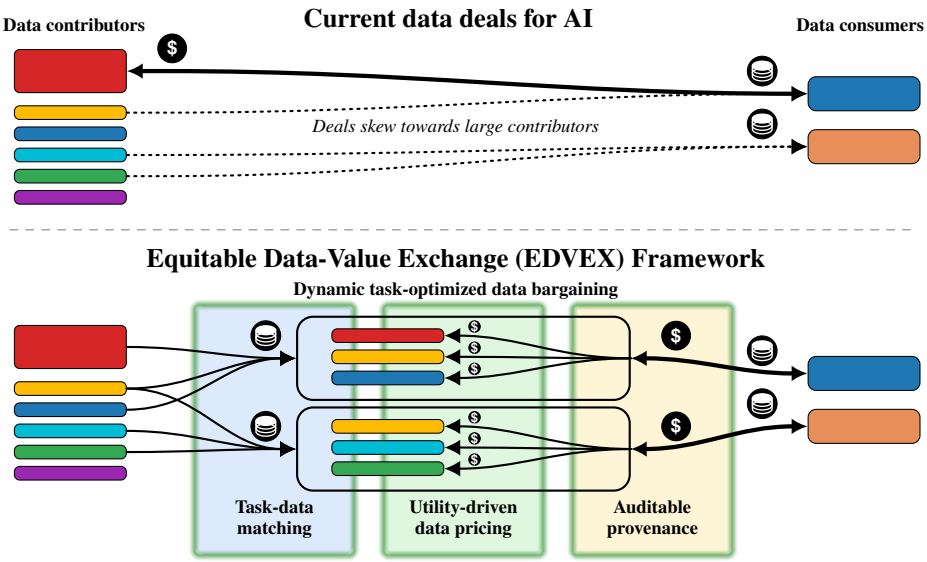

Figure 5: (**Top**): The current landscape of AI data deals is largely dominated by transactions with large-scale content holders (such as major publishers), lacking efficient price discovery mechanisms. Data from smaller players is often scraped *en masse* without compensation or simply overlooked. (**Bottom**): Our envisioned EDVEX Framework features efficient task-data matching, utility-driven data pricing, and auditable provenance to create a more efficient, equitable and transparent ecosystem.

training. To reduce the additional manual efforts for lineage tracking, the lineage tracking framework may be integrated with existing frameworks and logging metadata automatically. Interesting templates for such an architecture include cataloging projects such as Unity [43] and provenance-by-design protocol experiments such as Bittensor [44].

Open Problems for Tracking Lineage

Information requirements for lineage tracking. What specific information should be logged to enable effective lineage tracking? How granular should the metadata be regarding individual data creators, transformation processes, and intermediate outputs?
Balancing the metadata size and tracking accuracy. Given the potentially large amount of information needed for accurate lineage tracking, how can we design an efficient encoding mechanism? How should we navigate a trade-off between the metadata size and tracking accuracy?
Lower the barrier for tracking lineage. How can we design the software stack to minimize the manual effort? How can we efficiently ensure complete tracking with robust integrity protection?

Valuation. The current paradigm for data acquisition is often characterized by opaque, bilateral negotiations, frequently between large AI developers and data aggregators. This can systematically disadvantage not only smaller contributors but data buyers and may fail to reflect a dataset's true, context-dependent value [45]. This situation can lead to inefficient price discovery and potentially inequitable compensation [46], where significant data sources might be used without recompense or remain unutilized. The "Task-Data Matching" layer, by exploring the formation of dynamic, task-optimized data unions based on preliminary utility estimates, could lay some groundwork for more transparent and equitable valuation approaches. This layer would need to explore mechanisms for two core challenges: establishing a fair market price for an assembled data bundle and ensuring equitable revenue distribution among contributors within that union.

One avenue to explore is leveraging the task-specific utility estimates generated during the discovery phase as a foundation for more transparent price discovery. Instead of relying on bargaining power, the predicted performance uplift (e.g., loss reduction, accuracy gain) offered by a dynamically assembled data bundle could inform its market value. Several approaches could facilitate this. For instance, AI developers might bid for access to these task-optimized data bundles [47]. The utility estimates from the sandbox evaluation could serve as standard information for bidders, potentially fostering

more competitive and fair outcomes. As another example, the price of the bundle could be directly correlated with its estimated contribution to model performance [48]. Besides a one-time upfront payment, AI developers could also offer the union a share of the subsequent value generated by the model developed using their data (e.g., a percentage of revenue). This could align long-term incentives between AI developers and the data contributors. These mechanisms aim to create a more liquid and rational market for data, where price is more closely tied to demonstrable utility.

Once a price for the entire bundle is established, the proceeds can be shared among the contributing data sources. Crucially, for such dynamic data unions to be viable and encourage broad participation, equitable revenue sharing among contributors within the bundle is paramount [149]. If contributors do not perceive the allocation as fair, they may be reluctant to participate [49]. The implementation of equitable revenue sharing can draw inspiration from cooperative game theory concepts like the Shapley values [50], which determine the shares based on a participant's marginal contribution to the overall task-specific utility of the union. The sandbox mentioned above could be designed to provide useful signals to estimate the contribution of individual participants [51, 52].

Open Problems for Valuation

Efficient and reliable pre-acquisition estimation of data contribution. What evaluation processes should be conducted within the sandbox, and what specific information about candidate data sources must be made accessible for these evaluations, to enable the reliable and efficient estimation of their individual contributions *before* acquisition?
Understanding data's influence in complex and iterative AI development workflows. Modern AI development often involves intricate pipelines with multiple stages, diverse data types, varied training algorithms, and even iterative loops where models are trained on synthetic data generated by earlier model versions. How can we quantify the value contribution of an initial or intermediary dataset as it propagates and transforms through these sophisticated, multi-step processes?
Contribution to multi-faceted AI evaluation. How do we design data valuation mechanisms that reward contributions across multi-faceted performance metrics such as fairness and robustness?
Mitigating "gaming." Any data valuation system predicated on defined metrics is susceptible to "gaming," where contributors optimize for these metrics, potentially sacrificing genuine data quality [53]. How do we design valuation and market mechanisms that inherently reward genuinely useful data, while actively disincentivizing manipulative behaviors?
Addressing price erosion for highly substitutable data. How can valuation and market mechanisms be designed to prevent a "race to the bottom" for data contributions that are abundant and readily substitutable from numerous sources?

4 Related Work

The call for a more equitable data value chain is not new. Our proposal for an EDVEX builds upon, synthesizes, and extends several lines of existing work spanning conceptual frameworks, organizational models, technical implementations, and specific regulatory mechanisms.

Conceptual foundations: Data dignity, labor, and dividends. The philosophical underpinning of EDVEX resonates deeply with concepts like "Data Dignity" and "Data as Labor," notably championed by Jaron Lanier [54] and E. Glen Weyl [55]. These frameworks argue for recognizing the economic value of individual data contributions and advocate for systems where originators are compensated, aligning with EDVEX's core goals of traceable and equitable revenue-sharing. Similarly, proposals for "Data Dividends" (e.g., [56]), distributing profits from data back to contributors. While these ideas provide powerful normative grounding, EDVEX seeks to translate them into an actionable technical and economic protocol.

Organizational models for collective bargaining. To address the "asymmetric bargaining power" we identify, various organizational models have been proposed. Data Cooperatives enable members to pool data for collective negotiation and shared benefits (e.g., MIDATA in healthcare [34]). This goal of empowering data subjects through collective action is also central to Delacroix and Lawrence's exploration of bottom-up "data Trusts" [57], and is pursued through collective bargaining mechanisms in what Freedman discusses as data unions [58]. Building on this, this paper discusses mechanisms to facilitate the creation of dynamic data unions. These are specifically designed to empower data contributors—particularly those not part of established collectives—by enabling them to form

strategically around the requirements of specific ML tasks, thereby enhancing their capacity to influence ML model development and strengthen their bargaining power.

Online data marketplaces. Online data markets aim to enable data deals at scale, but they face significant hurdles regarding data pricing and data quality. Traditional data marketplaces [59, 60, 61, 62, 63, 64], for instance, typically utilize one-time upfront fees, query-based pricing, or subscriptions, which inadequately capture the context-dependent value of data, especially for machine learning applications. These platforms also offer limited information for potential buyers to evaluate dataset suitability, such as limited samples and metadata, making the process of finding suitable, high-quality data inefficient and uncertain. Recent efforts in the Web3 space [65, 66, 67, 44] have focused on creating decentralized data marketplaces, aimed to enhance transparency, introduce token-based compensation, and broaden participation by enabling more, often smaller, players to engage in data transactions. However, they often grapple with the same fundamental challenges of valuation and data quality, which the envisioned EDVEX addresses.

Regulatory frameworks. The evolving regulatory landscape increasingly recognizes the need for fairness and transparency in data handling, particularly with the rise of AI. Landmark regulations like the EU's General Data Protection Regulation (GDPR) [68] have established strong protections for personal data, emphasizing consent, data subject rights, and accountability. More recently, the EU Data Act [69] aims to ensure fairness in the allocation of data value in the digital economy, granting users (both individuals and businesses) greater rights to access and share data they co-generate and seeking to rebalance contractual power in data-sharing agreements. While these frameworks establish legal rights and obligations, this paper envisions a set of technical primitives that can help operationalize compliance and foster an ecosystem that embodies their spirit.

5 Counter Positions

Considering alternatives and the cost of inaction. The proposal for EDVEX stems from the need to address fundamental inefficiencies and inequities in the current data economy. However, it is crucial to consider alternative pathways to these goals and to understand the potential ramifications if these overarching issues remain unaddressed. Several strategies exist to improve the data value chain (see Section 4). However, they often present partial solutions or face limitations when aiming for systemic change. As noted in our discussion of regulatory frameworks (e.g., GDPR, EU Data Act), legal protections are advancing. However, such evolution, alongside current market practices of online data marketplaces (which, as discussed, struggle with robust valuation and discovery), may not sufficiently alter fundamental power imbalances or provide the utility-driven mechanisms EDVEX envisions for fair compensation and optimal data matching. Beyond current regulations, more direct governmental control over data access or value allocation could be pursued. However, this approach risks stifling innovation and may lack the agility to manage dynamic data markets effectively. Web3 initiatives and models like data cooperatives, trusts, and unions (detailed in Section 4) significantly advance transparency, collective bargaining, and contributor empowerment. EDVEX argues for more dynamic, task-optimized collaborations beyond static memberships, recognizing the task-dependent nature of data value. While collaborative ethical pledges by industry can be beneficial, they often lack robust enforcement and may not fully address the systemic representation and compensation issues for smaller data contributors.

Without addressing these fundamental challenges—opaque data valuation, limited access for smaller players, and inadequate compensation—the data economy will become increasingly concentrated, disadvantaging smaller innovators and stifling diverse AI development. These market inefficiencies will perpetuate the systematic undervaluation of data, while unfair practices diminish public trust and risk triggering restrictive regulatory responses [70]. This makes proactive, systemic solutions not just beneficial, but necessary.

Racing to the bottom. A significant consideration in designing equitable data deals is the potential for a "race to the bottom" in pricing, particularly for data that may appear abundant or easily substitutable. If many users can provide data suitable for a task, and only a subset is needed, market dynamics might indeed incentivize developers to select the lowest bidders, potentially devaluing the contributions of many. EDVEX's task-data matching layer aims to move beyond simple availability. By focusing on the marginal utility of data for specific tasks, and potentially identifying optimal combinations of diverse sources (as per [20, 30]), the system may value datasets not just on individual merit but on

their synergistic contribution. A data set that seems redundant in isolation might offer significant value (e.g., enhancing the representativeness of minority classes) when combined with others, thus resisting pure price-based selection. The concept of "dynamic, task-optimized data unions" is central here. While individual contributors of highly substitutable data might face downward price pressure, unions can provide collective bargaining power. These unions could establish minimum quality thresholds or value propositions for their pooled data, preventing a race to the bottom among their constituents for a given task requiring their specific collective offering. However, for data that is genuinely highly commoditized and where individual contributions offer little unique marginal utility even within optimal bundles, the risk of price depression remains a critical area for future research within the EDVEX framework (see Open Problems for Valuation).

Data for service. A pertinent consideration is the prevalent "data-for-service" model [71], where users receive non-monetary value through free access to services. This position paper does not inherently negate this exchange but seeks to bring transparency and fairness, particularly when data's utility extends beyond the immediate service provision. By making data's potential market value explicit through its valuation mechanisms, EDVEX could enable a clearer understanding of the "data" side of the bargain. This could facilitate scenarios where the value of a service is more consciously weighed against the value of data licensed for broader applications, potentially leading to new hybrid models where users are compensated for data uses that transcend their direct service experience.

Synthetic data. The increasing use of synthetic data in training AI models, and the emergence of iterative training loops where models generate data for further training ([72, 73, 74, 75, 76]), requires new considerations for a framework like EDVEX. It raises questions about the necessity for valuating human-generated data and the feasibility of tracing real data's value in these complex, recursive pipelines [77]. While synthetic data offers scalability and controllability ([31, 76]), human-generated data will likely remain crucial for grounding models in real-world distributions, nuances, and edge cases [78]. Synthetic data, especially if generated by models initially trained on other synthetic data, can suffer from a "model collapse" [10, 79]. In domains requiring direct interaction with the physical world—such as robotics ([80, 81]), autonomous driving ([82, 83]), and healthcare ([84, 85]), the need for authentic, real-world data for training, testing, and validation will persist and likely intensify [76]. Synthetic data can augment, but rarely fully replace, data from real-world sensors and interactions in these critical applications [31, 75]. Many valuable datasets are highly specialized, represent niche domains, or fall into the "long-tail" [86, 87]. Synthesizing high-quality, diverse data for these areas without sufficient initial real-world exemplars is extremely challenging [88, 89]. Overall, EDVEX's mechanisms for incentivizing the contribution of real-world datasets remain highly relevant.

It is crucial to also recognize that generating high-quality, diverse, and useful synthetic data is not a trivial or cost-free endeavor [73, 76]. Significant expertise, computational resources, and often sophisticated curation and filtering of initial seed data which may itself be real data are required [75, 77, 90]. EDVEX is agnostic to the origin of the data (real or synthetic) in principle; what matters is its utility, provenance, and the effort involved in its creation and curation. Thus, "data contributors" could indeed be entities (or individuals) who specialize in generating high-value synthetic datasets. We could assess the value of these synthetic contributions just as they would for real data.

6 Conclusion

The rapid advancement of artificial intelligence is inextricably linked to the availability and utilization of vast, diverse datasets. Yet, the current paradigms for data exchange are frequently marked by opacity, inefficiency, and an inequitable distribution of value that often disadvantages smaller contributors and hinders optimal data discovery. This position paper has explored EDVEX, a conceptual framework designed to address these foundational challenges. By considering integrated layers for task-data matching and discovery, auditable lineage tracking, and transparent, utility-driven valuation, this paper argues for a community effort to cultivate a data ecosystem that ensures bargaining symmetry, clear provenance, and efficient pricing for all participants.

Limitations. Our evidence base relies on publicly disclosed deals and public filings. However, many transactions are private or under NDA, so our dataset likely undercounts and is skewed toward larger, English-language, and U.S.-centric agreements. We therefore report patterns rather than exhaustive statistics. Confidentiality constraints preclude disclosure of non-public deal terms even when known, and we avoid speculation that could compromise sources. Market conditions have been evolving

rapidly in the last three years, limiting temporal generalization and our classification necessarily simplifies heterogeneous contracts. While we provide a transparent table, completeness cannot be assumed. As a position paper, we do not present an implementation or pilot. The feasibility and performance of key components remain open research problems.

Acknolwedgement

Ruoxi Jia and Feiyang Kang acknowledge support from the National Science Foundation through grants IIS-2312794, IIS-2313130, and OAC-2239622. Suqin Ge acknowledges support from the College of Science Dean's Discovery Fund at Virginia Tech. Jiachen T. Wang is supported by Apple's AI/ML PhD Fellowship, Princeton's Yan Huo *94 Graduate Fellowship and Princeton's Gordon Y.S. Wu Fellowship. Luis Oala thanks Bruno Sanguinetti and Freeman Lewin for engaging discussions during the review of the manuscript. The authors thank Alex Izydorczyk, Ithaka S+R and Freeman Lewin for providing resources to cross-reference public deal listings.

References

[1] Will openai ever make real money? *The Economist*, May 2025.

[2] The unbelievable scale of ai's pirated-books problem, 2025. The Atlantic, March 2025, accessed 2025-05-20.

[3] Reddit says it's made $203m so far licensing its data. https://techcrunch.com/2024/02/22/reddit-says-its-made-203m-so-far-licensing-its-data, 2024.

[4] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987, 2024.

[5] Royal free london nhs foundation trust and deepmind data sharing agreement, 2017. https://www.royalfree.nhs.uk/news/statement-re-deepmind-data-processing-agreement-during-testing-phase-streams-app.

[6] Pinterest privacy policy, 2025. Pinterest user data policy. https://policy.pinterest.com/en/privacy-policy.

[7] Openai faces early appeal in first ai copyright suit from coders, 2024. Bloomberg. https://www.datacenterknowledge.com/regulations/microsoft-github-openai-hit-with-code-copyright-lawsuit.

[8] Royal free - google deepmind trial failed to comply with data protection law, 2017. https://web.archive.org/web/20170705141936/https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/.

[9] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023.

[10] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[11] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.

[12] Google signs $60m reddit data licensing deal for ai training, 2024. https://www.cbsnews.com/news/google-reddit-60-million-deal-ai-training/.

[13] Google and stack overflow announce ai data partnership, 2024. https://www.wired.com/story/google-deal-stackoverflow-ai-giants-pay-for-data/.

[14] Ai (and other) companies: Quietly changing your terms of service could be unfair or deceptive, 2024. FTC, Feb 2024, accessed 2025-05-20.

[15] When the terms of service change to make way for a.i. training, 2024. NYT, June 2024, accessed 2025-05-20.

[16] Sherwin Rosen. The economics of superstars. *The American economic review*, 71(5):845–858, 1981.

[17] Chatgpt-maker openai signs deal with ap to license news stories, 2023. https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[18] Openai to start using news content from news corp. as part of a multiyear deal, 2024. https://apnews.com/article/openai-news-corp-a49144d381796df5729c746f52fbef19.

[19] Lingjiao Chen, Bilge Acun, Newsha Ardalani, Yifan Sun, Feiyang Kang, Hanrui Lyu, Yongchan Kwon, Ruoxi Jia, Carole-Jean Wu, Matei Zaharia, et al. Data acquisition: A new frontier in data-centric ai. *arXiv preprint arXiv:2311.13712*, 2023.

[20] Feiyang Kang, Hoang Anh Just, Anit Kumar Sahu, and Ruoxi Jia. Performance scaling via optimal transport: Enabling data selection from partially revealed sources. *Advances in Neural Information Processing Systems*, 36:61341–61363, 2023.

[21] Hoang A Just, I-Fan Chen, Feiyang Kang, Yuanzhi Zhang, Anit K Sahu, and Ruoxi Jia. Asr data selection from multiple sources: A practical approach on performance scaling. NeurIPS 2023 Workshop on Efficient Natural Language and Speech Processing ..., 2023.

[22] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Joan Giner-Miguelez, Pieter Gijsbers, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Michael Kuchnik, et al. Croissant: A metadata format for ml-ready datasets. *Advances in Neural Information Processing Systems*, 37:82133–82148, 2024.

[23] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[25] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[26] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

[27] Valeria Pais, Luis Oala, Daniele Faccio, and Marco Aversa. Autoguided online data curation for diffusion model training, 2025.

[28] Bernd Bischl, Giuseppe Casalicchio, Taniya Das, Matthias Feurer, Sebastian Fischer, Pieter Gijsbers, Subhaditya Mukherjee, Andreas C Müller, László Németh, Luis Oala, et al. Openml: Insights from 10 years and more than a thousand papers. *Patterns*, page 101317, 2025.

[29] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

[30] Feiyang Kang, Yifan Sun, Bingbing Wen, Si Chen, Dawn Song, Rafid Mahmood, and Ruoxi Jia. Autoscale: Automatic prediction of compute-optimal data composition for training llms. *arXiv preprint arXiv:2407.20177*, 2024.

[31] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[32] Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, et al. Datadecide: How to predict best pretraining data with small experiments. *arXiv preprint arXiv:2504.11393*, 2025.

[33] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 3363–3372, 2011.

[34] Midata. Accessed: 2025-05-21.

[35] Thibault Gisselbrecht, Sylvain Lamprier, and Patrick Gallinari. Dynamic data capture from social media streams: A contextual bandit approach. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 131–140, 2016.

[36] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[37] Luis Oala, Marco Aversa, Gabriel Nobis, Kurt Willis, Yoan Neuenschwander, Michèle Buck, Christian Matek, Jérôme Extermann, Enrico Pomarico, Wojciech Samek, et al. Data models for dataset drift controls in machine learning with optical images. *arXiv preprint arXiv:2211.02578*, 2022.

[38] Rasool Fakoor, Jonas Mueller, Zachary C Lipton, Pratik Chaudhari, and Alexander J Smola. Time-varying propensity score to bridge the gap between the past and present. *arXiv preprint arXiv:2210.01422*, 2022.

[39] Ghada Zamzmi, Kesavan Venkatesh, Brandon Nelson, Smriti Prathapan, Paul H Yi, Berkman Sahiner, and Jana G Delfino. Out-of-distribution detection and data drift monitoring using statistical process control. *arXiv preprint arXiv:2402.08088*, 2024.

[40] Amazon. Amazon SageMaker AI – Lineage Tracking Entities. `https://docs.aws.amazon.com/sagemaker/latest/dg/lineage-tracking-entities.html`. Developer Guide.

[41] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, et al. Developments in mlflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the fourth international workshop on data management for end-to-end machine learning*, pages 1–4, 2020.

[42] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. 2023.

[43] Unity catalog: Open, multimodal catalog for data & ai, 2025. `https://github.com/unitycatalog/unitycatalog`.

[44] Internet-scale neural networks, 2025. `https://github.com/opentensor/bittensor`.

[45] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[46] Ramesh Raskar, Praneeth Vepakomma, Tristan Swedish, and Aalekh Sharan. Data markets to support ai for all: Pricing, valuation and governance. *arXiv preprint arXiv:1905.06462*, 2019.

[47] Yingqi Gao, Jin Zhou, Hua Zhou, Yong Chen, and Xiaowu Dai. Learn then decide: A learning approach for designing data marketplaces. *arXiv preprint arXiv:2503.10773*, 2025.

[48] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 International Conference on Management of Data*, pages 1535–1552, 2019.

[49] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative ai at work. *The Quarterly Journal of Economics*, page qjae044, 2025.

[50] Lloyd S Shapley et al. A value for n-person games. 1953.

[51] Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024.

[52] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[53] Marilyn Strathern. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321, July 1997.

[54] Jaron Lanier. Jaron lanier fixes the internet. The New York Times. Accessed: 2025-05-21.

[55] Eric A Posner and E Glen Weyl. Radical markets: Uprooting capitalism and democracy for a just society. In *Radical Markets*. Princeton University Press, 2018.

[56] The Data Dividend Project. Accessed: 2025-05-21.

[57] Sylvie Delacroix and Neil D Lawrence. Bottom-up data trusts: Disturbing the 'one size fits all'approach to data governance. *International data privacy law*, 9(4):236–252, 2019.

[58] Eli Freedman. Data unions: The need for informational democracy. *Cal. L. Rev.*, 111:657, 2023.

[59] Amazon aws data exchange. `https://aws.amazon.com/data-exchange`, 2023.

[60] Databricks marketplace. `https://marketplace.databricks.com`, 2023.

[61] Narrative. `https://www.narrative.io`, 2023.

[62] Taus data marketplace. `https://datamarketplace.taus.net`, 2023.

[63] Gradient health. `https://gradienthealth.io`, 2023.

[64] Snowflake datamarketplace. `https://www.snowflake.com/en/data-cloud/marketplace/`, 2023.

[65] Ocean Protocol. Ocean protocol | tools for the new data economy.

[66] Masa Finance. Masa | decentralized zk-data network & marketplace.

[67] Sahara Labs Inc. Sahara | AI model network for decentralized knowledge exchange. `https://saharalabs.ai/`, 2024. Accessed: 2025-05-21.

[68] General data protection regulation. 2018.

[69] European Parliament and Council of the European Union. Regulation (eu) 2023/2854 of the european parliament and of the council of 13 december 2023 on harmonised rules on fair access to and use of data and amending regulation (eu) 2017/2394 and directive (eu) 2020/1828 (data act). Official Journal of the European Union, 2023.

[70] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. *Advances in Neural Information Processing Systems*, 37:108042–108087, 2024.

[71] Chris Anderson. *Free: The future of a radical price*. Random House, 2009.

[72] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

[73] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024.

[74] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

[75] Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

[76] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.

[77] Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. *arXiv preprint arXiv:2406.07515*, 2024.

[78] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Improving the scaling laws of synthetic data with deliberate practice. *arXiv preprint arXiv:2502.15588*, 2025.

[79] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.

[80] Ritvik Singh, Jingzhou Liu, Karl Van Wyk, Yu-Wei Chao, Jean-Francois Lafleche, Florian Shkurti, Nathan Ratliff, and Ankur Handa. Synthetica: Large scale synthetic data for robot perception. *arXiv preprint arXiv:2410.21153*, 2024.

[81] Celso M De Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*, 26(2):174–187, 2022.

[82] Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, et al. Synthetic datasets for autonomous driving: A survey. *IEEE Transactions on Intelligent Vehicles*, 9(1):1847–1864, 2023.

[83] NVIDIA Corp. NVIDIA Omniverse. `https://www.nvidia.com/en-us/omniverse/`, 2025. Accessed: 2025-05-23.

[84] Boris van Breugel, Tennison Liu, Dino Oglic, and Mihaela van der Schaar. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, 2(12):991–1004, 2024.

[85] Ghadeer O Ghosheh, Jin Li, and Tingting Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys*, 56(6):1–34, 2024.

[86] Osama Makansi, Özgün Çiçek, Yassine Marrakchi, and Thomas Brox. On exposing the challenging long tail in future prediction of traffic actors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13147–13157, 2021.

[87] Adam R Ferguson, Jessica L Nielson, Melissa H Cragin, Anita E Bandrowski, and Maryann E Martone. Big data from small data: data-sharing in the'long tail'of neuroscience. *Nature neuroscience*, 17(11):1442–1447, 2014.

[88] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 910–919. IEEE, 2024.

[89] P Bryan Heidorn. Shedding light on the dark data in the long tail of science. *Library trends*, 57(2):280–299, 2008.

[90] Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William A Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlaš, Ahmed Alaa, Adji Bousso Dieng, Natasha Noy, Vijay Janapa Reddi, James Zou, Praveen Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. DMLR: Data-centric machine learning research - past, present and future. *Journal of Data-centric Machine Learning Research*, 2024.

A Data Deals - Full Table

Data Receiver	Data Aggregator	Ref	Date	Type	$ Value	Codes
DeepMind	Moorfields Hospital	[91]	2016	Academic	Undisclosed	C
DeepMind	NHS	[92]	2017	Academic	Undisclosed	C
OpenAI	GitHub (Microsoft)	[128]	2018	UGC	Undisclosed	L
Adobe	Stock Contributors	[93]	2022	Images	Undisclosed	C,S
Various Licensees	X (formerly Twitter)	[107]	2023	UGC	2.5m/yr	C,R
OpenAI	Axel Springer	[103]	2023	News	20m+	C
Apple	Publishers	[94]	2023	News	Undisclosed	U
ElevenLabs	Voice Actors	[97]	2023	UGC	Undisclosed	C,S
IBM	NASA	[101]	2023	Images	Undisclosed	C
LG	Shutterstock	[98]	2023	Images	Undisclosed	C
Meta	Shutterstock	[99]	2023	Images	Undisclosed	C
Mubert	Musicians	[100]	2023	UGC	Undisclosed	C,S
NVIDIA	Getty Images	[102]	2023	Images	Undisclosed	C
OpenAI	Associated Press	[96]	2023	News	Undisclosed	C
OpenAI	Shutterstock	[106]	2023	Images	Undisclosed	C
OpenAI	StackOverflow	[138]	2023	UGC	Undisclosed	C
Perplexity	Multiple News Publishers	[104]	2023	News	Undisclosed	C,S,R
Runway	Getty Images	[105]	2023	Images	Undisclosed	C
Stability AI	AudioSparx	[95]	2023	UGC	Undisclosed	C
Stability AI	Getty Images	[152]	2023	Images	Undisclosed	L
Microsoft	Taylor & Francis / Informa	[108]	2024	Academic	10m	C
Undisclosed	HarperCollins	[116]	2024	Academic	2.5k/book	C,S
Undisclosed	Reuters	[109]	2024	News	22m	C
Amazon	Shutterstock	[118]	2024	Images	25-50m	C
Apple	Shutterstock	[119]	2024	Images	25-50m	C
Google	Shutterstock	[120]	2024	Images	25-50m	C
OpenAI	Shutterstock	[121]	2024	Images	25-50m	C
OpenAI	News Corp	[131]	2024	News	250m/5yr	C
Perplexity	Yelp	[134]	2024	UGC	25m	C
Large Tech Company	Wiley	[148]	2024	Academic	44m	C
Google	Reddit	[115]	2024	UGC	60m/yr	C
Undisclosed	Taylor & Francis / Informa	[140]	2024	Academic	65m	C
Undisclosed	Freepik	[117]	2024	Images	6m	C
Undisclosed	Tempus	[160]	2024	Health	72.8m	C,R
Google	StackOverflow	[114]	2024	UGC	Undisclosed	C
Meta	Reuters	[123]	2024	News	Undisclosed	C,U
Midjourney	Tumblr (Automattic)	[142]	2024	UGC	Undisclosed	C
Midjourney	Wordpress	[143]	2024	UGC	Undisclosed	C
Musical AI	Symphonic Distribution	[139]	2024	Audio	Undisclosed	C
NVIDIA	Shutterstock	[137]	2024	Images	Undisclosed	C
OpenAI	Dotdash Meredith	[113]	2024	News	Undisclosed	C
OpenAI	TIME	[141]	2024	News	Undisclosed	C
OpenAI	NYT	[124]	2024	News	Undisclosed	L
OpenAI	Reddit	[127]	2024	UGC	Undisclosed	C
OpenAI	Tumblr (Automattic)	[144]	2024	UGC	Undisclosed	C
OpenAI	Vox Media	[146]	2024	News	Undisclosed	C
OpenAI	Wordpress	[145]	2024	UGC	Undisclosed	C
OpenAI	Le Monde	[122]	2024	News	Undisclosed	C
OpenAI	Prisa Media	[129]	2024	News	Undisclosed	C
OpenAI	Financial Times	[125]	2024	News	Undisclosed	C
OpenAI	The Atlantic	[110]	2024	News	Undisclosed	C
OpenAI	Condé Nast	[112]	2024	News	Undisclosed	C
OpenAI	Axios	[157]	2024	News	Undisclosed	C
OpenAI	The Guardian	[154]	2024	News	Undisclosed	C
OpenAI	Schibsted	[130]	2024	News	Undisclosed	C
OpenAI	Future plc	[126]	2024	News	Undisclosed	C
OpenAI	Hearst Magazines	[132]	2024	News	Undisclosed	C
Potato	Wiley	[147]	2024	Academic	Undisclosed	C
ProRata AI	Multiple (500+) News Publishers	[135]	2024	News	Undisclosed	C,S
Undisclosed	Oxford University Press	[133]	2024	Academic	Undisclosed	U
Undisclosed	Cambridge University Press	[111]	2024	Academic	Undisclosed	U
Undisclosed	Sage	[136]	2024	Academic	Undisclosed	U
Amazon	New York Times	[150]	2025	News	Undisclosed	C
AWS	Wiley	[161]	2025	Academic	Undisclosed	C
Cohere	News/Media Alliance	[156]	2025	News	Undisclosed	L
Google	Associated Press	[153]	2025	News	Undisclosed	C
Mistral AI	Agence France-Presse	[155]	2025	News	Undisclosed	C
OpenEvidence	NEJM Group	[158]	2025	Academic	Undisclosed	C
Perplexity	Wiley	[162]	2025	Academic	Undisclosed	C
Pinterest	Pinterest Users	[159]	2025	Images	Undisclosed	C
ProRata	AAAS	[149]	2025	Academic	Undisclosed	C
Undisclosed	De Gruyter Brill	[151]	2025	Academic	Undisclosed	U
Undisclosed	DataSeeds AI (Zedge)	[163]	2025	Images	Undisclosed	C

Table 2: Joint table of data deals, in chronological order. Codes: C - deals confirmed by public sources, U - unclear deal status, L - litigation, S - revenue share, O - one-time payment, R - recurring payment.

Data Deal References

[91] Google deepmind and moorfields eye hospital partnership, 2016. Medical AI partnership. https://www.moorfields.nhs.uk/research/google-deepmind.

[92] Royal free london nhs foundation trust and deepmind data sharing agreement, 2017. https://www.royalfree.nhs.uk/news/statement-re-deepmind-data-processing-agreement-during-testing-phase-streams-app.

[93] Adobe firefly trained on adobe stock contributors' images, 2023. Adobe newsroom, Firefly FAQ. See https://www.adobe.com/sensei/generative-ai/firefly.html.

[94] Apple in talks to pay publishers up to $50 million for ai training, 2023. Reuters report. https://www.reuters.com/technology/apple-explores-ai-deals-with-news-publishers-new-york-times-2023-12-22/.

[95] Audiosparx signs licensing deal with stability ai, 2023. Accessed: 2024-07-03. See https://www.musicbusinessworldwide.com/stability-ai-launches-text-to-music-generator-trained-on-licensed-content-via-a-partnership-with-music-library-audiosparx/.

[96] Chatgpt-maker openai signs deal with ap to license news stories, 2023. https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[97] Elevenlabs voice marketplace for ai licensing, 2023. ElevenLabs marketplace. https://elevenlabs.io/voice-data-partnerships.

[98] Lg partners with shutterstock to advance ai for better life, 2023. https://www.shutterstock.com/blog/lg-partners-with-shutterstock-to-advance-ai-for-better-life.

[99] Meta licenses shutterstock music and images for ai training, 2023. Bloomberg Law, Meta MusicGen, Shutterstock press. https://www.prnewswire.com/news-releases/shutterstock-expands-long-standing-relationship-with-meta-301719769.html.

[100] Mubert licenses music from musicians for ai generation, 2023. Mubert docs. https://musically.com/2023/07/12/ai-music-startup-mubert-reaches-100m-tracks-milestone/.

[101] Nasa and ibm release open-source geospatial ai foundation model, 2023. https://www.earthdata.nasa.gov/news/nasa-ibm-collaborate-apply-ai-earth-science-data.

[102] Nvidia announces generative ai collaboration with getty images, 2023. https://blogs.nvidia.com/blog/generative-ai-getty-images/.

[103] Openai and axel springer sign content licensing deal, 2023. https://www.reuters.com/business/media-telecom/global-news-publisher-axel-springer-partners-with-openai-landmark-deal-2023-12-13/.

[104] Perplexity and publishers agree revenue-sharing model, 2023. https://pressgazette.co.uk/news/perplexity-publishers-revenue-sharing/.

[105] Runway strikes getty images deal ahead of gen-2 release, 2023. Accessed: 2024-07-03. See https://runwayml.com/news/runway-partners-with-getty-images.

[106] Shutterstock expands ai training data deal with openai, 2023. https://www.investopedia.com/shutterstock-expands-deal-with-openai-shares-rise-7559349.

[107] Twitter paid enterprise api access pricing revealed, 2023. https://mashable.com/article/twitter-elon-musk-paid-enterprise-api-access-pricing.

[108] Academic publisher strikes ai data deal with microsoft, 2024. Microsoft-Taylor & Francis deal. https://theconversation.com/an-academic-publisher-has-struck-an-ai-data-deal-with-microsoft-without-their-authors-knowledge-235203.

[109] Ai data licensing deals (magis), 2024. Reuters deal reference. https://magis.substack.com/p/ai-data-licensing-deals.

[110] The atlantic announces content partnership with openai, 2024. https://www.theatlantic.com/press-releases/archive/2024/05/atlantic-product-content-partnership-openai/678529/.

[111] Cambridge university press ai licensing opt-in, 2024. Cambridge AI licensing. https://www.thebookseller.com/news/cambridge-university-press--assessment-writes-to-20k-authors-for-ai-licensing-opt-in.

[112] Condé nast and openai partnership, 2024. https://www.condenast.com/news/conde-nast-openai-partnership.

[113] Dotdash meredith signs multi-year content deal with openai, 2024. Axios, May 29 2024. https://www.reuters.com/markets/deals/investopedia-owner-dotdash-meridith-signs-content-license-deal-with-openai-2024-05-07/.

[114] Google and stack overflow announce ai data partnership, 2024. https://www.wired.com/story/google-deal-stackoverflow-ai-giants-pay-for-data/.

[115] Google signs $60m reddit data licensing deal for ai training, 2024. https://www.cbsnews.com/news/google-reddit-60-million-deal-ai-training/.

[116] Harpercollins' ai deal will pay authors $2,500 per book, 2024. https://www.404media.co/harpercollins-ai-deal/.

[117] Inside big tech's underground race to buy ai training data, 2024. Freepik deal reference. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

[118] Inside big tech's underground race to buy ai training data (amazon), 2024. Amazon-Shutterstock deal. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

[119] Inside big tech's underground race to buy ai training data (apple), 2024. Apple-Shutterstock deal. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training_data-2024-04-05/.

[120] Inside big tech's underground race to buy ai training data (google), 2024. Google-Shutterstock deal. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

[121] Inside big tech's underground race to buy ai training data (openai), 2024. OpenAI-Shutterstock deal. https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/.

[122] Le monde signs ai partnership with openai, 2024. https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html.

[123] Meta signs deal with reuters to bring ai news to its platforms, 2024. https://www.axios.com/2024/10/25/meta-reuters-ai-news-facebook-instagram.

[124] Nyt v. openai: The times's about-face, 2024. Litigation reference. https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/.

[125] Openai and financial times announce partnership, 2024. https://aboutus.ft.com/press_release/openai.

[126] Openai and future partner on specialist content, 2024. https://openai.com/index/openai-and-future-partner-on-specialist-content/.

[127] Openai and reddit announce content partnership, 2024. Accessed: 2024-07-03. See https://www.theverge.com/2024/5/16/24158529/reddit-openai-chatgpt-api-access-advertising.

[128] Openai faces early appeal in first ai copyright suit from coders, 2024. Bloomberg. https://www.datacenterknowledge.com/regulations/microsoft-github-openai-hit-with-code-copyright-lawsuit.

[129] Openai partners with prisa media, 2024. https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/.

[130] Openai partners with schibsted media group, 2024. https://openai.com/index/openai-partners-with-schibsted-media-group/.

[131] Openai to start using news content from news corp. as part of a multiyear deal, 2024. https://apnews.com/article/openai-news-corp-a49144d381796df5729c746f52fbef19.

[132] Openai will bring hearst content to chatgpt, 2024. https://venturebeat.com/ai/openai-will-bring-cosmopolitan-publisher-hearsts-content-to-chatgpt.

[133] Oxford university press actively working with ai companies, 2024. LLM-Oxford deal. https://www.insidehighered.com/news/quick-takes/2024/08/05/oxford-university-press-actively-working-ai-companies.

[134] Perplexity chatbot yelp suggestions data ai, 2024. Perplexity-Yelp deal. https://www.theverge.com/2024/3/12/24098728/perplexity-chatbot-yelp-suggestions-data-ai.

[135] Prorata partners with dmg media, guardian, sky news and others, 2024. https://pressgazette.co.uk/platforms/prorata-ai-dmg-media-guardian-sky-news/.

[136] Sage confirms talks to license content to ai firms, 2024. LLMs-Sage deal. https://www.thebookseller.com/news/sage-confirms-it-is-in-talks-to-license-content-to-ai-firms.

[137] Shutterstock integrates generative ai across stock content service, 2024. https://www.businessinsider.com/shutterstock-integrated-gen-ai-stock-digital-photo-video-content-service-2024-12.

[138] Stack overflow signs data licensing agreement with openai, 2024. Accessed: 2024-07-03. See https://techcrunch.com/2024/05/06/stack-overflow-signs-deal-with-openai-to-supply-data-to-its-models/.

[139] Symphonic opens its catalogue for licensed ai training, 2024. https://musically.com/2024/08/21/symphonic-opens-its-catalogue-up-for-licensed-ai-training/.

[140] Taylor & francis ai deal sets worrying precedent, 2024. Undisclosed-Taylor & Francis deal. https://www.insidehighered.com/news/faculty-issues/research/2024/07/29/taylor-francis-ai-deal-sets-worrying-precedent.

[141] Time magazine strikes licensing deal with openai and perplexity, 2024. TIME press release, Jun 26 2024. https://www.reuters.com/technology/artificial-intelligence/openai-signs-multi-year-content-deal-with-time-magazine-2024-06-27/.

[142] Tumblr and wordpress to sell users' data to train ai tools, 2024. Midjourney-Tumblr deal. https://www.404media.co/tumblr-and-wordpress-to-sell-users-data-to-train-ai-tools/.

[143] Tumblr and wordpress to sell users' data to train ai tools (midjourney), 2024. Midjourney-Wordpress deal. https://www.404media.co/tumblr-and-wordpress-to-sell-users-data-to-train-ai-tools/.

[144] Tumblr and wordpress to sell users' data to train ai tools (openai), 2024. OpenAI-Tumblr deal. https://www.404media.co/tumblr-and-wordpress-to-sell-users-data-to-train-ai-tools/.

[145] Tumblr and wordpress to sell users' data to train ai tools (openai), 2024. OpenAI-Wordpress deal. https://www.404media.co/tumblr-and-wordpress-to-sell-users-data-to-train-ai-tools/.

[146] Vox media announces partnership with openai, 2024. Accessed: 2024-07-03. See https://www.theverge.com/2024/5/29/24167072/openai-content-copyright-vox-media-the-atlantic.

[147] Wiley creates ai partnership program, 2024. Potato-Wiley deal. https://www.publishersweekly.com/pw/by-topic/industry-news/industry-deals/article/96248-wiley-creates-ai-partnership-program.html.

[148] Wiley expects to make $44 million from ai partnership, 2024. Wiley deal reference. https://www.booksandpublishing.com.au/articles/2024/09/04/258068/wiley-expects-to-make-us44-million-from-ai-partnership-authors-unable-to-opt-out/.

[149] Aaas and prorata content licensing pilot, 2025. AAAS-ProRata deal. https://www.eurekalert.org/news-releases/1071967.

[150] Amazon signs ai licensing deal with the new york times, 2025. https://www.nytimes.com/2025/05/29/business/media/new-york-times-amazon-ai-licensing.html.

[151] De gruyter brill ai for authors, 2025. De Gruyter Brill AI deal. https://degruyterbrill.com/en/ai-for-authors/#:~:text=Why%20does%20De%20Gruyter%20Brill%20want%20to%20enter%20agreements%20with%20generative%20AI%20providers.

[152] Getty images sues stability ai over copyright infringement, 2025. https://www.dreyfus.fr/en/2025/02/19/getty-images-us-inc-and-others-v-stability-ai-ltd-2025-ewhc-38-ch-an-interesting-case-in-ai-and-intellectual-property-law/.

[153] Google signs deal with ap to deliver up-to-date news through gemini, 2025. https://www.ap.org/media-center/ap-in-the-news/2025/google-signs-deal-with-ap-to-deliver-up-to-date-news-through-its-gemini-ai-chatbot/.

[154] Guardian media group announces strategic partnership with openai, 2025. https://www.theguardian.com/gnm-press-office/2025/feb/14/guardian-media-group-announces-strategic-partnership-with-openai.

[155] Mistral signs deal with afp, 2025. https://ca.finance.yahoo.com/news/mistral-signs-deal-afp-offer-095158286.html.

[156] News/media alliance announces industry lawsuit, 2025. Litigation over news data. https://www.newsmediaalliance.org/news-media-alliance-announces-industry-lawsuit/.

[157] Openai will fund axios local newsrooms, 2025. https://www.axios.com/2025/01/15/open-ai-axios-local-newsrooms-funding-deal.

[158] Openevidence and nejm group announce partnership, 2025. OpenEvidence-NEJM deal. https://www.openevidence.com/announcements/openevidence-and-nejm.

[159] Pinterest privacy policy, 2025. Pinterest user data policy. https://policy.pinterest.com/en/privacy-policy.

[160] Tempus reports second quarter 2025 results; $3.7m in 1h 2024, 2025. https://www.tempus.com/news/tempus-reports-second-quarter-2025-results/?srsltid=AfmBOoqkGt9igPgNkP7xs-nB9mQHrrmXvWg_oQgfFKvzYnaVLSfzeYjK.

[161] Wiley announces collaboration with amazon web services, 2025. AWS-Wiley deal. https://newsroom.wiley.com/press-releases/press-release-details/2025/Wiley-Announces-Collaboration-With-Amazon-Web-Services-AWS-to-Integrate-Scientific-Content-Into-Life-Sciences-AI-Agents/default.aspx.

[162] Wiley expects to make $44 million from ai partnership (perplexity), 2025. Perplexity-Wiley deal. https://www.booksandpublishing.com.au/articles/2024/09/04/258068/wiley-expects-to-make-us44-million-from-ai-partnership-authors-unable-to-opt-out/.

[163] Zedge launches dataseeds ai content marketplace for ai training, 2025. https://www.stocktitan.net/news/ZDGE/zedge-launches-data-seeds-ai-a-content-marketplace-for-ai-training-yzg8lvti96ec.html.