Open Problems in AI Data Economics

Hamidah Oderinwale Anna Kazlaukas

hamidah@opendatalabs.xyz anna@opendatalabs.xyz

October 2025

Abstract

Recent research in machine learning has increasingly examined the economic impacts of new systems, yet little work has focused on the economics of AI development itself. This work lays the foundations for such research through three contributions. It formalizes data as a distinct factor of production alongside compute and labor, surveys emerging pricing and governance mechanisms, and identifies open problems in valuing and allocating data across the AI lifecycle. Drawing on historical cases of market standardization, it proposes concrete methods for measuring data's value, structuring ownership, and supporting its trade as an economic asset.

1 Introduction

Data remains the least understood of the three inputs to the still vaguely defined AI production function, even as scaling laws [1] highlight its role in driving frontier capabilities alongside compute and algorithms (architectures and optimization methods). To date, most economic research on AI emphasizes macro outcomes, such as the impacts of AI adoption on labor markets and productivity, while neglecting the production side. A related effort, "Open Problems in Technical AI Governance" [2], by Reuel et al. maps open questions in AI governance, including those concerning data governance.[1]

[1] While often used interchangeably, we refer to *machine learning* (ML) as the statistical and computational methods that underpin most current systems, and *artificial intelligence* (AI) as the broader goal of creating systems capable of performing cognitive tasks.

AI's economic footprint has become a growing focus for frontier labs developing widely used chat models.[2] While these initiatives reflect an institutional interest in quantifying AI's contribution to output and growth, they generally treat model capabilities as fixed inputs, leaving the internal production process opaque.

Missing from the research landscape, however, are microeconomic (endogenous) models that examine the production function itself, analyzing how data, compute, and skilled labor interact in forming economic models. For labs, these models would inform resource allocation; for policymakers, they would illuminate competition and governance issues; and for data contributors, they would clarify how supplying data generates value. Informally, labs already trade off compute for data by using open-weight models to generate synthetic datasets, as demonstrated by the Llama 3 series [7, 8].

Despite training on much of the public internet [9], frontier models still rely on less than 0.01% of the world's data. OpenMined and industry studies find that most organizations analyze only about 1% of what they collect—so-called "dark data." This gap highlights substantial opportunities for specialized, sovereign models [10, 11].

2 Markets and data valuation

To date, data pricing mechanisms [12] remain ad hoc and unsystematic in practice, negotiated on a case-by-case basis without frameworks to compare value across data types or use cases. Three main approaches have emerged: per-unit pricing, platform licensing, and service-based models.

2.1 Per-unit Pricing

We identify two models for per-unit licensing. The first is contractual: consider Microsoft's licensing deal with HarperCollins [13], licensing book titles at $5,000 each for three years of AI training rights, and image datasets priced from $0.01 to $0.25 per photo depending on quality and exclusivity. Licensing is distanced from model use, while transaction-based pricing links data value to consumption, compensating contributors in proportion to how often and how effectively their data powers model performance.

[2] Efforts such as Anthropic's Economic Index and Economic Futures Project [3, 4], Stripe's Economics of AI Fellowship [5], and OpenAI's GDPval benchmark [6].

The second, more nascent approach is transaction-based pricing, where data is priced per API call or query on model platforms or via inference providers. However, as models increasingly operate via local deployments where usage takes place off-platform and cannot be easily observed or measured, the evolution of pricing mechanisms that maintain usage-based compensation without centralized monitoring remains unclear.

2.2 Service-based Pricing

Service-based pricing sells the process, not the data, as buyers pay for annotation, curation, and cleaning that turn raw inputs into training-ready datasets. Scale AI exemplifies this model, offering large-scale annotation and data preparation as a service for model developers. In 2024, Google signed a $60 million annual agreement with Reddit for access to user-generated content [14]. That same year, Anthropic settled a $1.5 billion lawsuit over its use of pirated books to train Claude, a post-hoc example of data "pricing" through legal enforcement rather than market design [15].

2.3 Commissioning-based Pricing

Commissioning-based models represent a hybrid approach in which data is collected or curated for specific purposes, with the resulting outputs contributing to shared training or reference commons. Examples include Common Crawl, LAION, and research datasets for specialized assistants such as PubMed Central for biomedical research, arXiv for scientific reasoning, and legal case repositories for jurisprudence. What distinguishes these models is their emphasis on continuous feedback and maintenance. Data improves through use: as datasets are applied, gaps and errors are revealed, enabling correction and refinement over time [16]. Because these systems (e.g., RAG [17]) are knowledge-based, their capabilities remain tightly coupled to the quality and evolution of the underlying datasets (i.e., via stewardship).

Table 1: Data Pricing Mechanisms

Pricing Mechanism	Description	Example(s)	Pricing
Per-Unit / Query Pricing	Data sold per item or unit with fixed or tiered prices.	Claude API: $0.25 input to $75 per million tokens output; Scale AI annotation pricing per record.	$0.01–$5,000+ per unit
Platform Licensing	Subscription or tiered fee for data access bundled with tools.	ChatGPT Pro: $20–$200/month; Claude AI: $20–$200/month; Snowflake Marketplace: 750+ providers.	$20–$250/month
Service-Based Models	Managed data pipelines or annotation services.	Scale AI, Claude Code, Cursor IDE (Pro: $20/month).	Per task or tiered fees
Major Licensing Deals	Large-scale data agreements.	Google–Reddit: $60M/yr [14]; Shutterstock: $25–$50M [18]; HarperCollins: $5,000/book [19].	$10M–$60M+ per deal

3 Data, its properties, and historical asset precedents

Unlike traditional commodities, data has distinct economic properties. It is non-rivalrous, since reuse does not diminish supply, and only partially excludable, as access can be restricted but copies are easily made. However, contamination (where data quality degrades over time or is poisoned) or overuse (when widespread training reduces data's competitive value) can create practical rivalry effects, reducing datasets' value for future users. Data is difficult to measure dynamically, poorly standardized, and traded in new, underdeveloped markets where value is set through bespoke deals rather than price benchmarks. These differences help explain why data markets behave differently from traditional commodity markets [20].

Furthermore, data's value is highly context-dependent. A dataset that meaningfully improves performance during pre-training may add little during fine-tuning, and vice versa. Some datasets exhibit combinatorial effects, becoming far more valuable when combined with others, especially across domains. Their contribution depends not only on content but also on the systems they train, the stage of the pipeline where they are used, and the

complements such as storage, versioning systems, and file formats that make them usable.

We can look to history to understand how traditional commodities such as natural resources (like oil, timber, and land) and regulated intangibles (like spectrum) only became tradable once institutions made them measurable, standardized, and investable through contracts [21]. Grain markets once suffered from chronic quality disputes and fragmented regional pricing until USDA grading and futures contracts brought standardization [22]. Oil markets moved from chaotic local trade to globally benchmarked commodities through API gravity standards and price indexes like West Texas Intermediate. Similar transformations shaped energy, spectrum, timber, and real estate.

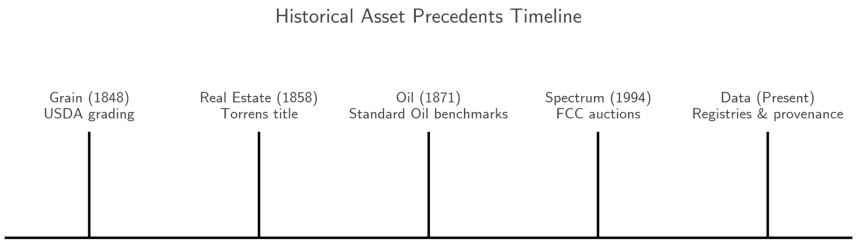

Figure 1: Timeline of key assets and the institutional innovations that structured their markets.

Table 2: Historical Asset Market Development

Asset Class	**Mechanism**	**Developer**	**Date(s)**	**Outcome**	**Properties**
Agriculture	Grain futures and warehouse receipts: standardized grading and storage contracts	Chicago Board of Trade (CBOT)	1848	Stabilized food supply chains and enabled hedging against harvest volatility	Rivalrous and perishable; value secured through storage and timing.

Real Estate	Torrens Title System: standardized, verifiable ownership records	Robert Torrens; South Australian Government	1858–Present	Made land financeable and enabled securitization through REITs	Scarce and immobile; value stems from exclusive control and location.
Oil	Futures contracts and spot benchmarks (West Texas Intermediate, Brent Crude): standardized delivery and pricing	Joseph Leiter; Chicago Board of Trade (CBOT); later NYMEX, OPEC	1870s–1900s	Shifted oil from local commodity to globally benchmarked and traded resource	Fungible and storable; uniform value across contexts.
Energy	Levelized Cost of Electricity (LCOE): standardized lifetime cost metric for power generation	Peter F. Drucker (precursor); IEA; U.S. EIA	1960s	Created predictable cost curves that enabled long-term infrastructure investment	Consumable and rivalrous; value depends on real-time delivery.
Timber	Forest Stewardship Council (FSC): international sustainability certification	Tim Synnott; Forest Stewardship Council	1993	Shifted industry from extractive to renewable, managed resource base	Rivalrous and renewable; value tied to stewardship cycles.
Spectrum	Spectrum auctions: market-based allocation of frequency bands	Evan Kwerel; John McMillan; Federal Communications Commission	1994	Enabled efficient allocation of a scarce intangible resource powering mobile networks	Rivalrous and intangible; value enforced through licensed access.

Data	Emerging proposals for standardized data valuation, licensing, and registries	Diane Coyle; Laura Veldkamp; Jian Pei; OECD; Vana	2020s–	Transforming data from opportunistically traded byproduct to recognized capital asset	Non-rivalrous, replicable, and context-dependent; value shaped by access rights, quality, and complementarities.

4 Revisiting economic theory

Because AI depends so heavily on data, economic theory must expand to show how data interacts with other inputs. Earlier technological shifts pushed economists to look beyond capital and labor; today, production models must recognize data as a core factor of production.

Early growth models focused narrowly on tangible inputs. The 1946 Roy Harrod–Evsey Domar model assumed output depended only on fixed ratios of physical capital and labor.

In 1956, Robert Solow and Trevor Swan introduced technological progress as a key driver of sustained growth, with later economists like Gary Becker and Robert Lucas extending this framework in the 1960s–1980s to include human capital.

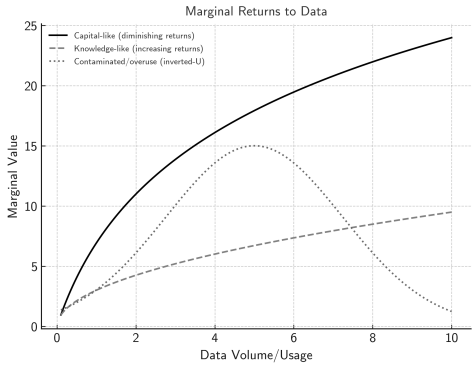

Figure 2: Three stylized models for data's contribution to AI production: diminishing returns (capital-like), sustained or increasing returns with quality, and inverted-U under contamination or overuse.

In 1990, Paul Romer treated knowledge itself as capital, emphasizing increasing returns and spillovers. Elinor Ostrom showed how communities govern common-pool resources without central control.

Open questions:

1. If production depends on fixed factors, where does data fit? Should it be treated like capital, like labor, or as a new input altogether with different properties?

2. Does data behave like physical capital with diminishing returns, or like human capital where quality shifts can sustain long-term performance?

3. Under what conditions does data exhibit increasing rather than diminishing returns, and how can spillovers across firms or domains be measured?

4. Given that data can be copied freely but its value depends on context, how can we design governance mechanisms that prevent underinvestment, misallocation, or loss of provenance?

5 The state of research on data economics

5.1 Measuring data's contribution to model performance

Farboodi and Veldkamp [23] model data as information with endogenous value, explaining barter-like exchanges of services for user data. Pei [12] surveys data pricing and cost structures; Coyle [24] applies real-options theory to data investments. These works establish that data's value is context-dependent and endogenous—shaped by how it is used, combined, and deployed. Yet existing research has not resolved fundamental measurement problems: data's cost structure (cheap to copy, costly to verify) remains poorly reflected in pricing mechanisms.

5.2 Market structure and information asymmetries

Hooker's essay "The Hardware Lottery" [25] highlights complementarities between hardware and data that create bottlenecks to scientific and algorithmic progress. Santesteban and Longpre [26] document data monopolies; Agarwal et al. [27] analyze auction-based markets and exclusivity arrangements. Bergemann et al. [28] show that uncertainty around data value distorts pricing and may hinder scaling of fine-tuning-as-a-service models. These studies reveal that information asymmetries and complementarities between data and infrastructure create structural barriers to efficient market formation.

This research landscape identifies the gaps but does not prescribe solutions. Moving from diagnosis to design requires establishing institutional foundations—standards, metrics, and governance mechanisms—that allow data to function as a tradable factor of production. The Four Pillars framework that follows positions the essential infrastructure and open research problems that should anchor this emerging field.

6 Foundations for AI data economics: Four pillars

Treating data as a factor of production requires establishing four foundational pillars: how data is classified and measured, how its value evolves across production stages, which models best capture data's role in AI output, and what units and metrics make that role tractable. Together, these

pillars define the institutional infrastructure necessary for data to function as a tradable asset.

6.1 How should data be classified and measured as an asset?

Markets function when goods are defined, measurable, and trackable. Energy has kilowatt hours, farmland has acres, and stocks have shares. Data lacks equivalent units. Its intrinsic value depends on context and is difficult to measure directly. Prices align with intrinsic value only when markets are transparent. For exchange trading, markets need standardized products, known supply, and transaction systems [29].

Key questions:

1. How should data's future value be accounted for when its current contribution cannot be directly measured?
2. What models could make the value of data in barter-like transactions measurable and comparable across contexts?
3. How can ownership and transfer rights be verified when value often lies in combinations and transformations?

   Emerging approaches include watermarking and fingerprinting of model outputs (e.g., DeepMind's SynthID [30, 31]), cryptographic provenance and audit trails [32, 33], and standardized content attribution frameworks such as the Content Provenance and Authenticity standard (C2PA) [34, 35].
4. Can informational content be valued separately from the competitive advantage of exclusive access?
5. When model use is opaque for open-weight deployments and local inference, what enforcement mechanisms can replace metering?
   - Cryptographic proof of use, such as TEEs or merkle proofs of data lineage [36, 37].
   - Periodic auditing with statistical sampling [38, 39].
   - Hybrid models using lump-sum licensing with spot checks [40, 35].
   - Pricing tiers based on actor reputation for repeat use [41, 42].

6. What standardized could make data value comparable across sectors (e.g., flow-based metrics like kilowatt hours for energy or MHz-pop in spectrum auctions)?

6.2 How does data's value evolve across the ML supply chain?

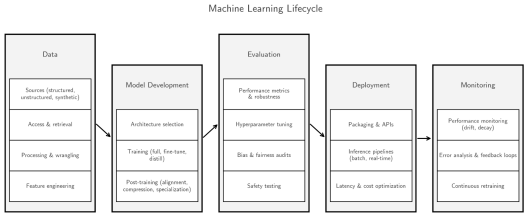

Figure 3: Diagram of the machine learning lifecycle.

At each stage, data interacts differently with compute and human expertise. Pre-training relies on large-scale but hard-to-value corpora; post-training requires curated task data; inference generates real-time, user-specific data. RLHF [43] and test-time training [44] shift the economic importance of data types.

Key questions:

1. Pipeline value paradox: Why does identical data have different marginal productivity across stages?

2. Continuous learning dilemma: Do depreciation models fit "living" assets that update continuously?

3. Lifecycle arbitrage: If stage values differ predictably, why do markets not equalize returns intertemporally?

4. Sparse allocation problem: In Mixture-of-Experts, only subsets of parameters see each example. Recent work on FlexOlmo [45] explores independent experts trained on closed datasets and combined through domain-aware routing, with experts toggled at inference. How should data be priced when its contribution can be switched on or off, and when value depends on which other experts are active?

6.3 What Models Best Capture AI Production and Data's Role Within It?

AI production maps data, compute, and human expertise to outputs such as accuracy and capabilities. Data can be reused across models; compute needs scale superlinearly; human expertise serves as both input and QA. Quality often dominates quantity; curated datasets can outperform massive noisy ones. Synthetic data [46] changes substitution dynamics.

Key questions:

1. Curation paradox: Do highly curated datasets exhibit increasing returns to quality?
2. Synthetic inheritance: When synthetic data is derived from proprietary inputs, how much is value creation versus transfer?
3. Substitution impossibility: Can quality substitute for compute, or are they strict complements?
4. Non-rivalrous monopoly: How can firms sustain advantages using a non-consumable resource?
5. Data-limited frontier: If compute becomes abundant (e.g., via accelerator optimization [47]) but high-quality data is scarce, how should models reflect a data bottleneck?

6.4 What Units and Metrics Make Data Economics Tractable?

Valuation begins with units. Vana's Data Capital Locked (DCL) [48] proposes estimating pooled data value by individual contributions. But marginal value is context-dependent: what is the contribution of the 1,000th fitness record versus the first? How should prices reflect diminishing or non-linear returns? How do combined datasets create emergent value?

Key questions:

1. Cardinal vs. ordinal: Can data be measured absolutely, or only ranked within use cases?
2. Temporal decay: How do we create stable units when value changes with model and method advances?

3. Network effects: If value scales non-linearly with network size, how do units capture exponential relationships?

4. Composite valuation: When value arises from the whole, what is the measurable unit?

5. 7 Building the Field of Data Economics

Treating data as capital means building tools, standards, and institutions to measure value, assign ownership, and govern use. Engineers know which datasets improve models but lack valuation methods; economists model markets but often miss technical realities. Closing the gap requires embedded economists in labs and shared frameworks. Legal scholars can clarify ownership and rights; engineers can build quality, privacy, and provenance systems; creators and practitioners can ground models in real production; interdisciplinary teams can translate theory into deployable tools.

8 Conclusion

This work lays the groundwork for studying AI data economics by identifying open questions and situating them within the existing literature. As an economic asset, information remains nascent, and the sparse research to date leaves fundamental problems in measurement, valuation, and governance unresolved. The goal here is to chart these problems, focusing on data's contribution to the AI production function and the evolution of the surrounding ecosystem.

Data differs from capital, labor, and commodities: it is reusable, combinatorial, and context-dependent. Without shared standards, we risk market concentration, opaque pricing, and the exclusion of smaller actors. Advancing data economics requires concrete frameworks that capture marginal value, verify provenance and ownership, and model interactions with compute and labor. The task ahead is to classify, measure, and govern data so that it functions as a tradable and accountable factor of production.

References

[1] 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". In: *arXiv preprint arXiv:2203.15556* (2022).

[2] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. *Open Problems in Technical AI Governance*. 2025. arXiv: 2407.14981 [cs.CY]. URL: https://arxiv.org/abs/2407.14981.

[3] Anthropic. *Economic Index Geography*. Accessed: 2025-10-07. 2024. URL: https://www.anthropic.com/research/economic-index-geography.

[4] Anthropic. *Economic Futures Project*. Accessed: 2025-10-07. 2024. URL: https://www.anthropic.com/economic-futures.

[5] Stripe. *Economics of AI Fellowship*. Accessed: 2025-10-07. 2024. URL: https://stripe.events/fellowship.

[6] OpenAI. *GDPval: Evaluating Model Capabilities on Economically Valuable Tasks*. Accessed: 2025-10-07. 2024. URL: https://openai.com/index/gdpval/.

[7] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[8] NVIDIA Developer Blog. *Creating Synthetic Data Using Llama 3.1 405B*. Accessed October 2025. 2024. URL: https://developer.nvidia.com/blog/creating-synthetic-data-using-llama-3-1-405b/.

[9] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. *Will we run out of data? Limits of LLM scaling based on human-generated data*. Accessed: 2025-10-07. 2024. URL: https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

[10] OpenMined. *AI is trained and evaluated on less than 0.01% of the world's data*. Accessed: 2025-10-07. 2025. URL: https://openmined.org/.

[11] Wikipedia contributors. *Dark data*. Accessed: 2025-10-07. 2025. URL: https://en.wikipedia.org/wiki/Dark_data.

[12] Jian Pei. "A survey on data pricing: from economics to data science". In: *IEEE Transactions on Knowledge and Data Engineering* 34.10 (2020), pp. 4586–4608.

[13] Jay Peters. *Microsoft reportedly made an AI training deal with HarperCollins*. The Verge, Accessed: 2025-10-07. 2024. URL: https://www.theverge.com/2024/11/19/24300893/microsoft-ai-training-deal-harpercollins-report.

[14] Reuters. *Reddit in AI content licensing deal with Google, sources say*. Accessed: 2025-10-07. 2024. URL: https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/.

[15] CNN. *Anthropic AI settlement with authors over Claude*. Accessed: 2025-10-07. 2025. URL: https://www.cnn.com/2025/09/05/business/anthropic-ai-settlement-authors-claude.

[16] Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. *Language Modeling with Editable External Knowledge*. 2024. arXiv: 2406.11830 [cs.CL]. URL: https://arxiv.org/abs/2406.11830.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

[18] VentureBeat. *Apple's $25–50 million Shutterstock deal highlights fierce competition for AI training data*. Accessed: 2025-10-07. 2024. URL: https://venturebeat.com/ai/apples-25-50-million-shutterstock-deal-highlights-fierce-competition-for-ai-training-data/.

[19] The Guardian. *HarperCollins to allow tech firms to use books to train AI models*. Accessed: 2025-10-07. 2024. URL: https://www.theguardian.com/books/2024/nov/19/harpercollins-tech-firms-books-train-ai-models-nonfiction-artificial-intelligence.

[20] John Baffes and Peter Nagle. *Commodity markets: evolution, challenges, and policies*. World Bank Publications, 2022.

[21] John McMillan. "Selling Spectrum Rights". In: *Journal of Economic Perspectives* 8.3 (Sept. 1994), pp. 145–162. DOI: 10.1257/jep.8.3.145. URL: https://www.aeaweb.org/articles?id=10.1257/jep.8.3.145.

[22] Aziz Elbehri. "The changing face of the US grain system: differentiation and identity preservation trends". In: (2007).

[23] Maryam Farboodi and Laura Veldkamp. "Long-run growth of financial data technology". In: *American Economic Review* 111.8 (2021), pp. 2485–2523.

[24] Diane Coyle. "Measuring the value of data: challenges and approaches". In: *International Journal of the Economics of Business* 31.2 (2024), pp. 261–280.

[25] Sara Hooker. *The Hardware Lottery*. 2020. arXiv: 2009.06489 [cs.CY]. URL: https://arxiv.org/abs/2009.06489.

[26] Cristian Santesteban and Shayne Longpre. "How big data confers market power to big tech: Leveraging the perspective of data science". en. In: *Antitrust Bull.* 65.3 (Sept. 2020), pp. 459–485.

[27] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. *A Marketplace for Data: An Algorithmic Solution*. 2019. arXiv: 1805.08125 [cs.GT]. URL: https://arxiv.org/abs/1805.08125.

[28] Dirk Bergemann and Alessandro Bonatti. "Data, Competition, and Digital Platforms". In: *American Economic Review* 114.8 (Aug. 2024), pp. 2553–2595. DOI: 10.1257/aer.20230478. URL: https://ideas.repec.org/a/aea/aecrev/v114y2024i8p2553-95.html.

[29] Ertem Nusret Tas, István András Seres, Yinuo Zhang, Márk Melczer, Mahimna Kelkar, Joseph Bonneau, and Valeria Nikolaenko. "Atomic and fair data exchange via blockchain". In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. Salt Lake City UT USA: ACM, Dec. 2024, pp. 3227–3241.

[30] Jose Fernandez et al. "Data watermarking for machine learning datasets: Techniques and challenges". In: *ACM Computing Surveys* (2023).

[31] Google DeepMind. *SynthID: Identifying AI-generated images and audio*. `https://deepmind.google/discover/blog/synthid-identifying-ai-generated-images-and-audio/`. 2023.

[32] Thomas Hardjono and Alex Pentland. "Data provenance and integrity using blockchain and trusted computing". In: *IEEE Security & Privacy* (2019).

[33] Franziska Boenisch et al. "Provenance-enabled audit trails for machine learning systems". In: *NeurIPS Workshop on Trustworthy Machine Learning*. 2021.

[34] C2PA Consortium. *Content Provenance and Authenticity (C2PA) Specification*. `https://c2pa.org/specifications/specifications/`. 2022.

[35] Bill Rosenblatt et al. "Data Licensing in the Age of AI". In: *Journal of Intellectual Property Law & Practice* (2024).

[36] Raymond Cheng et al. "Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts". In: *IEEE European Symposium on Security and Privacy*. 2019.

[37] Bohan Zhang et al. "Proof of Learning: Definitions and Practice". In: *arXiv preprint arXiv:2203.08575* (2022).

[38] Ari Juels and Burton Kaliski. "PoRs: Proofs of Retrievability for Large Files". In: *ACM CCS*. 2007.

[39] Yixin Chen et al. "Auditing the Auditors: Ensuring Data Integrity in Machine Learning Systems". In: *arXiv preprint arXiv:2304.06721* (2023).

[40] Avi Goldfarb and Daniel Trefler. "Generative AI and the Future of Work: How Pricing Models Shape Innovation Incentives". In: *NBER Working Paper No. 31702* (2023).

[41] Paul Resnick, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara. "Reputation systems". In: *Communications of the ACM* 43.12 (2000), pp. 45–48.

[42] Patrick Bolton, Xavier Freixas, and Joel Shapiro. "Credit reports, reputation, and lending". In: *Review of Financial Studies* 18.4 (2005), pp. 1253–1300.

[43] Nathan Lambert. *Reinforcement Learning from Human Feedback*. 2025. arXiv: 2504.12501 [cs.LG]. URL: https://arxiv.org/abs/2504.12501.

[44] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. *s1: Simple test-time scaling*. 2025. arXiv: 2501.19393 [cs.CL]. URL: https://arxiv.org/abs/2501.19393.

[45] Weijia Shi, Akshita Bhagia, Kevin Farhat, Niklas Muennighoff, Pete Walsh, Jacob Morrison, Dustin Schwenk, Shayne Longpre, Jake Poznanski, Allyson Ettinger, Daogao Liu, Margaret Li, Dirk Groeneveld, Mike Lewis, Wen-tau Yih, Luca Soldaini, Kyle Lo, Noah A. Smith, Luke Zettlemoyer, Pang Wei Koh, Hannaneh Hajishirzi, Ali Farhadi, and Sewon Min. *FlexOlmo: Open Language Models for Flexible Data Use*. 2025. arXiv: 2507.07024 [cs.CL]. URL: https://arxiv.org/abs/2507.07024.

[46] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. *Best Practices and Lessons Learned on Synthetic Data*. 2024. arXiv: 2404.07503 [cs.CL]. URL: https://arxiv.org/abs/2404.07503.

[47] Bowen Peng, Jeffrey Quesnelle, and Diederik P. Kingma. *DeMo: Decoupled Momentum Optimization*. 2024. arXiv: 2411.19870 [cs.LG]. URL: https://arxiv.org/abs/2411.19870.

[48] Vana. *Data Capital Locked (DCL) — docs.vana.org*. https://docs.vana.org/docs/data-capital-locked-dcl. [Accessed 09-10-2025]. 2025.

References

[1] 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". In: *arXiv preprint arXiv:2203.15556* (2022).

[2] Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, David Bau, Paul Bricman, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. *Open Problems in Technical AI Governance*. 2025. arXiv: 2407.14981 [cs.CY]. URL: https://arxiv.org/abs/2407.14981.

[3] Anthropic. *Economic Index Geography*. Accessed: 2025-10-07. 2024. URL: https://www.anthropic.com/research/economic-index-geography.

[4] Anthropic. *Economic Futures Project*. Accessed: 2025-10-07. 2024. URL: https://www.anthropic.com/economic-futures.

[5] Stripe. *Economics of AI Fellowship*. Accessed: 2025-10-07. 2024. URL: https://stripe.events/fellowship.

[6] OpenAI. *GDPval: Evaluating Model Capabilities on Economically Valuable Tasks*. Accessed: 2025-10-07. 2024. URL: https://openai.com/index/gdpval/.

[7] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[8] NVIDIA Developer Blog. *Creating Synthetic Data Using Llama 3.1 405B*. Accessed October 2025. 2024. URL: https://developer.nvidia.com/blog/creating-synthetic-data-using-llama-3-1-405b/.

[9] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. *Will we run out of data? Limits of LLM scaling based on human-generated data*. Accessed: 2025-10-07. 2024. URL: https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

[10] OpenMined. *AI is trained and evaluated on less than 0.01% of the world's data*. Accessed: 2025-10-07. 2025. URL: https://openmined.org/.

[11] Wikipedia contributors. *Dark data*. Accessed: 2025-10-07. 2025. URL: https://en.wikipedia.org/wiki/Dark_data.

[12] Jian Pei. "A survey on data pricing: from economics to data science". In: *IEEE Transactions on Knowledge and Data Engineering* 34.10 (2020), pp. 4586–4608.

[13] Jay Peters. *Microsoft reportedly made an AI training deal with HarperCollins*. The Verge, Accessed: 2025-10-07. 2024. URL: https://www.theverge.com/2024/11/19/24300893/microsoft-ai-training-deal-harpercollins-report.

[14] Reuters. *Reddit in AI content licensing deal with Google, sources say*. Accessed: 2025-10-07. 2024. URL: https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/.

[15] CNN. *Anthropic AI settlement with authors over Claude*. Accessed: 2025-10-07. 2025. URL: https://www.cnn.com/2025/09/05/business/anthropic-ai-settlement-authors-claude.

[16] Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. *Language Modeling with Editable External Knowledge*. 2024. arXiv: 2406.11830 [cs.CL]. URL: https://arxiv.org/abs/2406.11830.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

[18] VentureBeat. *Apple's $25–50 million Shutterstock deal highlights fierce competition for AI training data*. Accessed: 2025-10-07. 2024. URL: https://venturebeat.com/ai/apples-25-50-million-shutterstock-deal-highlights-fierce-competition-for-ai-training-data/.

[19] The Guardian. *HarperCollins to allow tech firms to use books to train AI models*. Accessed: 2025-10-07. 2024. URL: https://www.theguardian.com/books/2024/nov/19/harpercollins-tech-firms-books-train-ai-models-nonfiction-artificial-intelligence.

[20] John Baffes and Peter Nagle. *Commodity markets: evolution, challenges, and policies*. World Bank Publications, 2022.

[21] John McMillan. "Selling Spectrum Rights". In: *Journal of Economic Perspectives* 8.3 (Sept. 1994), pp. 145–162. DOI: 10.1257/jep.8.3.145. URL: https://www.aeaweb.org/articles?id=10.1257/jep.8.3.145.

[22] Aziz Elbehri. "The changing face of the US grain system: differentiation and identity preservation trends". In: (2007).

[23] Maryam Farboodi and Laura Veldkamp. "Long-run growth of financial data technology". In: *American Economic Review* 111.8 (2021), pp. 2485–2523.

[24] Diane Coyle. "Measuring the value of data: challenges and approaches". In: *International Journal of the Economics of Business* 31.2 (2024), pp. 261–280.

[25] Sara Hooker. *The Hardware Lottery*. 2020. arXiv: 2009.06489 [cs.CY]. URL: https://arxiv.org/abs/2009.06489.

[26] Cristian Santesteban and Shayne Longpre. "How big data confers market power to big tech: Leveraging the perspective of data science". en. In: *Antitrust Bull.* 65.3 (Sept. 2020), pp. 459–485.

[27] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. *A Marketplace for Data: An Algorithmic Solution*. 2019. arXiv: 1805.08125 [cs.GT]. URL: https://arxiv.org/abs/1805.08125.

[28] Dirk Bergemann and Alessandro Bonatti. "Data, Competition, and Digital Platforms". In: *American Economic Review* 114.8 (Aug. 2024), pp. 2553–2595. DOI: 10.1257/aer.20230478. URL: https://ideas.repec.org/a/aea/aecrev/v114y2024i8p2553-95.html.

[29] Ertem Nusret Tas, István András Seres, Yinuo Zhang, Márk Melczer, Mahimna Kelkar, Joseph Bonneau, and Valeria Nikolaenko. "Atomic and fair data exchange via blockchain". In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. Salt Lake City UT USA: ACM, Dec. 2024, pp. 3227–3241.

[30] Jose Fernandez et al. "Data watermarking for machine learning datasets: Techniques and challenges". In: *ACM Computing Surveys* (2023).

[31] Google DeepMind. *SynthID: Identifying AI-generated images and audio*. `https://deepmind.google/discover/blog/synthid-identifying-ai-generated-images-and-audio/`. 2023.

[32] Thomas Hardjono and Alex Pentland. "Data provenance and integrity using blockchain and trusted computing". In: *IEEE Security & Privacy* (2019).

[33] Franziska Boenisch et al. "Provenance-enabled audit trails for machine learning systems". In: *NeurIPS Workshop on Trustworthy Machine Learning*. 2021.

[34] C2PA Consortium. *Content Provenance and Authenticity (C2PA) Specification*. `https://c2pa.org/specifications/specifications/`. 2022.

[35] Bill Rosenblatt et al. "Data Licensing in the Age of AI". In: *Journal of Intellectual Property Law & Practice* (2024).

[36] Raymond Cheng et al. "Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts". In: *IEEE European Symposium on Security and Privacy*. 2019.

[37] Bohan Zhang et al. "Proof of Learning: Definitions and Practice". In: *arXiv preprint arXiv:2203.08575* (2022).

[38] Ari Juels and Burton Kaliski. "PoRs: Proofs of Retrievability for Large Files". In: *ACM CCS*. 2007.

[39] Yixin Chen et al. "Auditing the Auditors: Ensuring Data Integrity in Machine Learning Systems". In: *arXiv preprint arXiv:2304.06721* (2023).

[40] Avi Goldfarb and Daniel Trefler. "Generative AI and the Future of Work: How Pricing Models Shape Innovation Incentives". In: *NBER Working Paper No. 31702* (2023).

[41] Paul Resnick, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara. "Reputation systems". In: *Communications of the ACM* 43.12 (2000), pp. 45–48.

[42] Patrick Bolton, Xavier Freixas, and Joel Shapiro. "Credit reports, reputation, and lending". In: *Review of Financial Studies* 18.4 (2005), pp. 1253–1300.

[43] Nathan Lambert. *Reinforcement Learning from Human Feedback*. 2025. arXiv: 2504.12501 [cs.LG]. URL: https://arxiv.org/abs/2504.12501.

[44] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. *s1: Simple test-time scaling*. 2025. arXiv: 2501.19393 [cs.CL]. URL: https://arxiv.org/abs/2501.19393.

[45] Weijia Shi, Akshita Bhagia, Kevin Farhat, Niklas Muennighoff, Pete Walsh, Jacob Morrison, Dustin Schwenk, Shayne Longpre, Jake Poznanski, Allyson Ettinger, Daogao Liu, Margaret Li, Dirk Groeneveld, Mike Lewis, Wen-tau Yih, Luca Soldaini, Kyle Lo, Noah A. Smith, Luke Zettlemoyer, Pang Wei Koh, Hannaneh Hajishirzi, Ali Farhadi, and Sewon Min. *FlexOlmo: Open Language Models for Flexible Data Use*. 2025. arXiv: 2507.07024 [cs.CL]. URL: https://arxiv.org/abs/2507.07024.

[46] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. *Best Practices and Lessons Learned on Synthetic Data*. 2024. arXiv: 2404.07503 [cs.CL]. URL: https://arxiv.org/abs/2404.07503.

[47] Bowen Peng, Jeffrey Quesnelle, and Diederik P. Kingma. *DeMo: Decoupled Momentum Optimization*. 2024. arXiv: 2411.19870 [cs.LG]. URL: https://arxiv.org/abs/2411.19870.

[48] Vana. *Data Capital Locked (DCL) — docs.vana.org*. https://docs.vana.org/docs/data-capital-locked-dcl. [Accessed 09-10-2025]. 2025.