Paying for Information, Not for Clicks: A Block-Level Publisher Content Marketplace for AI Answers

Markus Mobius[1], Luca Braghieri[6], Sarah Eichmeyer[6], Daniel G. Goldstein[1], Nicole Immorlica[1], Oliver Kiss[2], Ro'ee Levy[7], Brendan Lucier[1], Ishan Mahajan[3], Jingyi Qiu[4], Tanya Rosenblat[4], David M. Rothschild[1], Andras Sari[5], Aleksandrs Slivkins[1], and Colin Watson[3]

[1] *Microsoft Research*
[2] *Spreadmonitor*
[3] *Microsoft*
[4] *University of Michigan*
[5] *Central European University*
[6] *University of Bocconi*
[7] *Tel Aviv University*

October 10, 2026

Working paper

Abstract

We propose a publisher content marketplace that allocates payments for information used in AI-generated answers. The unit of attribution is a "lego block": a distinct fact or piece of commentary that may appear in articles from multiple publishers. A simple news-reading model connects three payment rules: equal sharing among all publishers reporting on a block, rewarding the publisher who reported a block first, and rewarding indispensable contributions. Each can be represented within a Shapley-value framework.

We evaluate the framework using answers to 1,974 search-engine queries about a heavily covered news episode. When attribution considers only the articles retrieved for an individual query, most cited blocks appear unique, and the rules produce nearly identical allocations. A *global index* covering the full news corpus reveals overlap across publishers, making the choice of payment rule consequential. Premium publishers, defined as established outlets with substantial newsgathering operations, receive 42% of allocated value under equal sharing, compared with 46–50% under a first-publication rule, depending on the treatment of publication timestamps. These findings highlight the importance of a shared global index that records which publishers provide each block and when they published it. Attribution restricted to the sources retrieved for an individual answer lacks the information needed to distinguish these payment policies.

Keywords: content marketplaces, attribution, Shapley value, generative search, news economics, market design.
JEL: D47, L82, L86, O34.

*Mobius is first author. All other authors are listed alphabetically. The lego decomposition approach was developed by Braghieri, Eichmeyer, Kiss, Levy, Mobius, Rosenblat and Sari. Qiu collaborated with this group on adapting the decomposition into the lego implementation. Goldstein, Immorlica, Lucier, Mobius, Rosenblat, Rothschild and Slivkins developed the theory framework for measuring attribution. Mahajan and Watson helped frame the problem. All errors are our own.

1 Introduction

Historically, publishers of online content were paid in clicks. An intermediary (e.g., search engine or a social feed) showed a headline, the reader clicked through to the publisher's page, and the publisher monetized the visit. The intermediary profited, too, receiving a payment for the referral to the publisher.

Generative search breaks this arrangement. When an assistant answers a question directly, it summarizes the publisher's reporting while providing the publisher almost no traffic. Chapekis and Lieb (2025) find that users click through on 8% of searches carrying AI summaries (compared to 15% without), and click a link *inside* the summary on just 1%. Yet the underlying content is more valuable to the intermediary than ever, as a grounded answer cannot be generated without it.

This shift has created a market-design problem that platforms are actively attempting to solve: finding alternative channels through which to value content and pay publishers. For example, Cloudflare's Pay-Per-Crawl marketplace (Allen and Newton, 2025) and Amazon's AWS AI traffic monetization system (Kayabali, 2026), unofficially dubbed the "AWS AI Content Marketplace," allow publishers to charge AI crawlers for content access. While these systems allow publishers to set prices, they do not directly assist with price discovery or value determination. In February 2026, Microsoft announced a Publisher Content Marketplace to pay publishers on "delivered value" through usage-based reporting, launching with partners such as the Associated Press, Condé Nast, Vox Media, and the USA TODAY Company (Frank, 2026). While this announcement establishes the *ambition* of paying for the content that's actually used, it leaves the underlying *mechanism* open.

Why is it challenging to monetize content directly? The difficulty begins with the unit of account. Media consumption has unbundled over two decades. News used to be consumed in newspapers consisting of collections of articles. With the rise of the internet, that was unbundled into individual articles, and then rebundled through news aggregators. Music underwent a similar shift from album to track to multi-source playlist, re-pricing the underlying media at every step.

With AI, there is an additional unbundling of content, from full coherent pieces like news articles, to their components – facts or comments. The market for publisher content must adjust its pricing model accordingly: if a marketplace's unit of account remains the article or webpage, it measures something that is no longer consumed. This suggests monetizing at a more granular level, but facts and comments are difficult to price. They are *non-rival* (one reader's consumption does not deplete the fact or comment) (Jones and Tonetti, 2020) and *non-excludable* (once published, anyone can legally re-report it) (Arrow, 1962). In practice, non-excludability yields rampant *duplication*. In our data, a cited news fact appears in the reporting of a mean of 22.4 distinct publishers: one filed first, while others carried the same fact through independent reporting or aggregation.

This paper makes four core contributions:

Contribution 1: A unit of account for news and answer generation. We propose a fine-grained, interpretable unit of account for news content. News reporting can be broken down into atomic *"lego" blocks*: *news* blocks record state-changing actions in the world, while *commentary* blocks record stances, opinions, or context regarding those actions.

This unit offers two primary advantages over existing approaches. It is *explainable*: blocks are short declarative sentences that publishers can audit directly without relying on obscure model saliency scores or counterfactual re-generations[1]. It is also *comparable*: a single comment or fact, reported by many distinct publishers, is a single block, reducing redundancy.

Contribution 2: A system for indexing articles and answering queries with attribution. We describe a system that constructs a coherent index of lego blocks from a corpus of news articles. Each crawled article is decomposed by a language model into constituent blocks and arranged into a tree that represents their relationships. Identical blocks across publishers are clustered and deduplicated into a shared, tagged index that records every atomic claim, its publishers, and their respective timestamps. This generates a shared global index of lego blocks that represent the content of multiple articles.

We demonstrate that building this representation at index time simplifies the query response process at inference time, allowing queries to be settled via straightforward table lookups. Furthermore, generating answers from deduplicated blocks significantly reduces token overhead by inserting shared facts or comments into the model's context window once rather than repeatedly. Finally, since the index retains publisher information for each block, provenance is retained and credit attribution can be calculated whenever an answer is generated. In contrast, an index constructed at inference time retains only knowledge of a single source of a given block but not the collection or reporting times. Thus computing attribution at inference time, e.g., by perturbing or inspecting retrieved context post-query (Cohen-Wang et al., 2024; Qi et al., 2024; Nematov et al., 2025), cannot distinguish an originator from a copier because that global context was discarded prior to query execution, limiting its ability to implement sophisticated attribution rules.

Contribution 3: A unifying economic theory of reader value and credit attribution. A primary motivation for credit attribution is determining how to appropriately divide payments among publishers for their content. Once news reports map to lego blocks, the payment problem simplifies: for every block used in an answer, the index identifies who reported it and when. This information can be used as input to a *payment rule* that compensates publishers.

We describe a flexible space of payment rules that can be used. We evaluate three intuitive payment rules drawn from existing literature:

[1]Counterfactual re-generation techniques estimate source influence at inference time by repeatedly removing retrieved documents from the context window, re-running the language model, and measuring changes in the output text (see, e.g., Cohen-Wang et al., 2024; Qi et al., 2024).

1. Equal division among all publishers that carried the block (EQUAL), mirroring music-streaming pro-rata models (Datta et al., 2018);

2. Exclusive payment to the originator (PROVENANCE), addressing the under-rewarding of original scoops (Cagé et al., 2020);

3. Payment based on indispensability (COMPLETER), following the default posture of data valuation via the Shapley value (Ghorbani and Zou, 2019; Jia et al., 2019).

We show that all three rules can be interpreted as Shapley values for different formulations of a content-creation game. Imagine a reader who reviews grounded articles in sequence and registers new blocks as she reads while ignoring blocks she already encountered. Each publisher is compensated for the marginal utility added when the reader reads their article. Setting the reading order to *random* and reader utility to *linear* in the facts discovered yields EQUAL; setting the order to *chronological* yields PROVENANCE; and holding the order random while pushing utility toward an *all-or-nothing* model yields COMPLETER. This framework provides a structured foundation for evaluating payment mechanisms, mapping them as points along two parameterized economic dimensions: the *order dial* (whether speed deserves reward) and the *utility dial* (how heavily completeness is valued).[2]

Contribution 4: Empirical findings and index-time necessity. We implement and test our system against a broad corpus of news articles related to the Iran/US war in the summer of 2026. We find that basing attribution on a global index rather than a local one has a large empirical impact. Against a global index, only 27.2% of cited blocks have a single publisher, and cited blocks have an average of 22.4% publishers. We find that different payment rules separate clearly, shifting premium publishers' revenue share from 41.87% under EQUAL to 46.3–49.6% under PROVENANCE. Against a local index, however, 86.2% of cited blocks appear unique within the retrieval set (with a mean coverage of 1.23), causing every payment rule in the family to collapse to the same payout within one percentage point. This suggests that credit attribution is most meaningful when evaluated against a global pre-constructed index: attribution rules matter only when duplication is visible, and duplication is only visible globally. We also find that the EQUAL rule raises the relative value of commentary over original news reporting, due to commentary blocks being less duplicated than news blocks. This suggests that a credit attribution system based on provenance can meaningfully help to reward original reporting by premium publishers in a content marketplace. On the other hand, we find that the PROVENANCE rule concentrates payments toward a small number of large originators, away from the tail of smaller publishers.

[2]The model is also generative: setting both dials to *chronological* and *all-or-nothing* reveals an unstudied fourth rule that compensates the *last* publisher needed to complete an answer. We view its emergence as a by-product of the two dials rather than a proposal.

Roadmap The paper proceeds as follows. Section 2 reviews related literature. Section 3 describes the index, its construction, and its use in answer generation. Section 4 describes attribution scoring options and develops the unified Shapley mechanism. Section 5 describes our datasets and empirical results, and Section 6 concludes.

2 Related literature

This paper proposes a market-design approach to news markets in an AI-centric world. As such, the research relates to prior work on the economics of news, referral systems, product unbundling, information goods, data valuation, generative attribution, and relevant legal frameworks. We discuss each in turn.

Economics of News Reporting. A long tradition studies how news market structures shape reporting incentives (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2010; Gentzkow et al., 2015; Prat and Strömberg, 2013). Closest to our work is Cagé et al. (2020), who document that a significant share of online news is aggregated or copied rather than originally reported, leaving original reporting under-rewarded as copies appear almost instantaneously. Their finding provides the empirical motivation for this paper. Our framework measures this phenomenon at the level of individual facts rather than entire articles: we show that a cited news block is carried by a mean of 22.4 publishers but originated by just one, and our order dial explicitly dictates whether that head start is monetized. Related work by Cagé (2020) and Angelucci et al. (2024) explores the broader implications for media competition and public information diets.

Aggregators and the Unbundling of Media. How news content is monetized has been studied extensively through the lens of search engines and news aggregators, focusing on the traditional referral-traffic bargain (Athey et al., 2021; Chiou and Tucker, 2017; Calzada and Gil, 2020; Aguiar and Waldfogel, 2021). Generative search changes this dynamic fundamentally: rather than reducing referral traffic, it eliminates it (Chapekis and Lieb, 2025; Newman et al., 2025). Consequently, the market-design challenge shifts from analyzing bargaining power to establishing a precise metering standard. Unbundling content from published bundles to atomic components is a well-established economic trajectory (Bakos and Brynjolfsson, 1999), seen most clearly in the music industry's transition from albums to tracks (Datta et al., 2018). Music streaming faced a nearly identical challenge—distributing a shared subscription pool across rights holders—and settled on pro-rata and user-centric allocation rules that closely mirror some of the credit attribution rules we consider.

Economics of Information Goods. Underpinning this structure is the economics of non-rival information goods (Jones and Tonetti, 2020), where market transactions are famously constrained by Arrow's Information Paradox: a buyer cannot evaluate information without consuming it, yet

once consumed, cannot "un-see" it (Arrow, 1962; Bergemann and Bonatti, 2019; Bergemann et al., 2022). A block-level index addresses this paradox by making content inspectable in granular units that can be priced without exposing the full article text. More broadly, our approach aligns with the normative framework of treating data as labor in digital economies (Arrieta-Ibarra et al., 2018).

Technical Data Valuation. Technically, our framework connects to the literature on data valuation in machine learning. The Shapley value has emerged as the standard theoretical instrument (Ghorbani and Zou, 2019; Jia et al., 2019; Wang et al., 2025), often estimated via influence functions or datamodels due to computational constraints (Koh and Liang, 2017; Ilyas et al., 2022; Park et al., 2023), with applications extending to royalty schemes for generative platforms (Wang et al., 2024). Crucially, however, that literature values data for *training*; we value content for *serving* (retrieval augmented generation).

Attribution vs. Settlement in Generative AI. Generative attribution literature serves as the primary contrast class for Section C. Current technical approaches evaluate citations at inference time (Rashkin et al., 2021; Bohnet et al., 2022; Gao et al., 2023; Liu et al., 2023) or attribute generated text via context ablation, perturbation, or internal model introspection (Cohen-Wang et al., 2024; Qi et al., 2024; Goldshmidt and Horovicz, 2024; Nematov et al., 2025). Our claim is not that these methods fail at their intended task, but rather that their task is answer *verification*, not financial *settlement*. Because inference-time methods are constrained to a query's local retrieval set by construction, they are structurally incapable of executing content compensation policies—a limitation we formalize in Section C and quantify in Section 5.2. One exception is ProRata.ai (Business Wire, 2024), which employs attribution methods on generated text to assign fractional credit and compensation to different data sources. Their approach is notably post-generation, applying to output text and external to the generation process, whereas our proposed marketplace attributes credit for data retrieved during generation itself. Complementary systems literature on passage granularity (Chen et al., 2023) and corpus deduplication (Lee et al., 2022) provides additional engineering support for our block-level representation.

Legal and Policy Context. Finally, legal frameworks surrounding fair use, content licensing, and copyright consent are evolving rapidly (Lemley and Casey, 2021; Henderson et al., 2023; Sag, 2023; Longpre et al., 2024). We take no stance on legal doctrines, but emphasize that regardless of the legal theory adopted, almost every proposed remedy requires a robust, fine-grained measurement layer that does not currently exist.

3 The Lego Index

This section describes the global index of lego blocks, which serves as the settlement layer of our proposed marketplace. The index represents a corpus of articles in a form that makes answer retrieval and economic attribution both feasible and computationally efficient. The index is a relational table of atomic claims, each tagged with the outlets that reported it and their respective publication timestamps. Storing these two fields directly, rather than recomputing them per query, reduces attribution to a table lookup.

3.1 Decomposing an article into a block tree

Our index decomposes the text of each article in a corpus into a tree of atomic *blocks*. Blocks fall into two distinct functional categories:

News blocks record an *action*: a verifiable state change in the world brought about by individuals, organizations, or nature (e.g., a vote, a strike, an attack, or a hurricane). They document who acted, what occurred, when, and where. Because these facts form the skeleton of news articles, they are frequently duplicated across outlets.

Commentary blocks record an opinion or reaction (typically through a quote for non-opinion pieces), piece of context, or analytical stance *about* a news action. They tend to be less duplicated across publishers and are often optional: the reporting of the underlying event can stand without them.

Separating news from commentary isolates the two economically distinct goods a publisher sells. Facts tend to be reported by multiple publishers, so provenance-based attribution methods significantly impact them. Commentary, on the other hand, is often unique to a single publisher, causing it to play a different role in attribution. Section 5.2 investigates the differential impact of attribution rules on facts and commentary.

Extracted blocks are organized into a structured *tree* rather than an unorganized bag of claims. Each commentary block is attached to the specific news block it references. Similarly, a news block may logically follow a preceding news block that precipitated it (e.g., a ceasefire hanging below the strikes that preceded it). The decomposition of an article into a tree of blocks takes several language-model passes and runs as soon as the search index crawler discovers a new article. Figure 1 illustrates a worked example.

We maintain this tree structure for two reasons. First, it is essential for clustering and deduplication, as we describe in Section 3.2. Second, the tree representation serves a vital role during answer generation, as we discuss in Section 3.4. Furthermore, although we do not address it here, attribution methods for certain types of content, such as academic papers, might wish to provide attribution to blocks that are predecessors of cited ones.

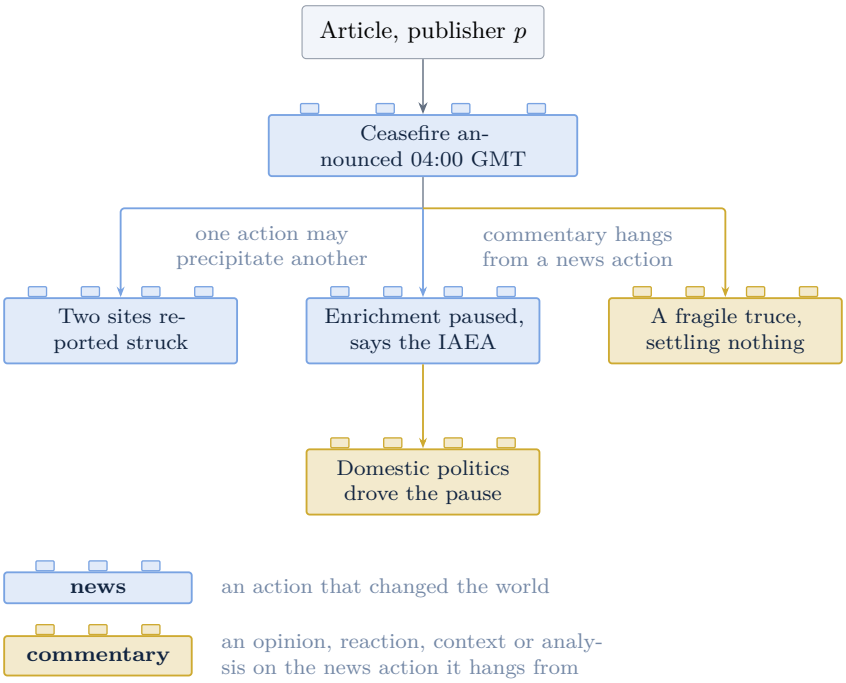


Figure 1: A block tree. One article decomposed into news and commentary blocks.

3.2 Clustering across publishers

Once a block trees have been constructed for each individual article, they are combined into a unified index. Blocks expressing equivalent semantic content are clustered *across* and *within* publishers, mapping a block reported by twenty outlets to a single index row with twenty associated publishers rather than twenty separate entries. This clustering and deduplication process relies heavily on the tree structure constructed for each article. In isolation, a commentary sentence like *"This decision was a mistake"* cannot be meaningfully matched across publishers: two such statements in different outlets might refer to entirely different news actions. The parent node establishes the necessary context, so deduplication clustering necessarily evaluates a commentary block together with its parent news block.

These two steps—transforming an article into a lego tree of blocks and then deduplicating those blocks—transform a list of URLs into the *global index*. Each cluster row records every contributing publisher along with its publication timestamp. Figure 2 traces this ingestion process: shared claims stack additional publisher tags on existing rows, while novel claims append new rows.

3.3 Properties, scope, and limitations

The global index is a representation of a news corpus, optimized for attribution. As such, the index is bounded by the reach of the corpus upon which it is based. Furthermore, the index measures *coverage*, not *veracity*. It records that an outlet published a claim, not whether the

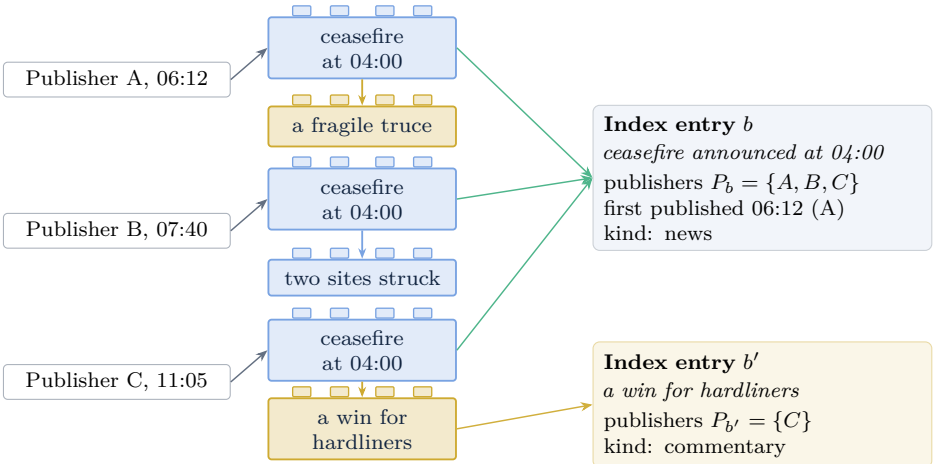


Figure 2: Building the index. Each publisher contributes a block tree; blocks from separate trees that express the same claim collapse into a single index row. The row records who carries the claim and who filed first.

claim is factual, and no payment rule in this paper conditions payout on truthfulness. In practice, this distinction is less restrictive than it might appear. Across the broader media ecosystem, fabricated news represents a tiny fraction of total consumption (Allen et al., 2020; Grinberg et al., 2019; Guess et al., 2020), meaning most news blocks record real-world events. Commentary blocks, by contrast, frequently quote external statements where accurate reporting and factual accuracy diverge. The index faithfully records the reporting of the claim; distributing payments based on verified truth would require an auxiliary verification signal outside the index itself. In any case, accurately reporting a false statement made by someone else is not fabrication, and the commentary block is a correct record of the reporting either way.

The global index can easily be made inspectable: publishers could directly inspect their indexed lego blocks, verify timestamp records, and audit the set of co-reporting publishers. We expect this transparency to be a prerequisite for any practical, functional content marketplace, as publishers will retain the ability to audit metrics upon which their payments are based.

3.4 Answering queries and attributing credit using the index

When a query arrives, a search engine retrieves a grounding set of relevant articles. The choice of grounding set is optimized by the retrieval system and is not under the direct control of our system; it may or may not include all applicable articles, but is designed to include sufficient coverage to address the given query. In a conventional retrieval-augmented generation (RAG) pipeline, the raw text of those articles is placed directly into the model's context window, and the model generates an answer accompanied by URL citations. In our pipeline, the retrieved

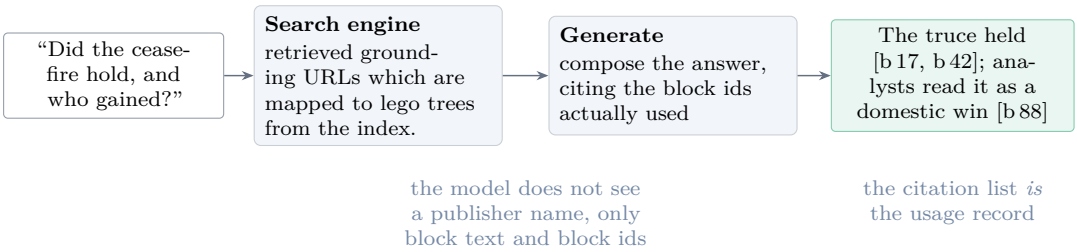


Figure 3: Answering Queries. A query arrives, deduplicated block trees rather than raw article text are placed into context, and the model cites specific block identifiers in its response.

articles are instead represented by their pre-computed lego block trees. The model is prompted to synthesize its response using these deduplicated blocks and to explicitly cite the unique block identifiers (b) it relied upon. Figure 3 illustrates this query pipeline.

Deduplicating facts before insertion into the context yields a computational benefit by reducing context length. Constructing a global index therefore does not inflate query-time generation costs; rather, by removing redundant facts prior to context assembly, it makes answer generation cheaper. We evaluate these savings empirically in Section 5.2.

Finally, credit for the answer must be divided among publishers. Fortunately, this credit attribution does not impose a significant extra calculation cost on the answer generation process. Because the model explicitly outputs cited block identifiers, and each block is associated with a list of articles to which it can traced, attribution requires no additional post-hoc LLM calls or context perturbations. Furthermore, publishers reporting a block can receive attribution even if they are not returned by grounding as the global index contains a record of their report alongside the block itself. Settlement reduces to a lookup against the indexed articles. What remains is to chose how attribution is calculated, which is detailed in Section 4.

3.5 Scope and Timing

We emphasize two aspects of credit attribution in our system. First, it is evaluated against the global index of all publishers, as opposed to a local index consisting only of publishers represented in the grounding set returned for a given query. Second, attribution of credit for each block is computed at index time, while the global index is being built; not at inference time, at the point where a query is to be answered.

Attribution performed at index time and at a global scope requires an up-front cost to build and maintain the index, but saves on per-query computation costs and provides a more holistic view of the content landscape from which to evaluate provenance. This distinguishes our method from various inference-time attribution approaches, such as asking the model to cite its sources (Gao et al., 2023; Bohnet et al., 2022); perturbing the context and re-running, as in ContextCite (Cohen-Wang et al., 2024) or Shapley over retrieved documents (Nematov et al., 2025); and reading the model's internals (Qi et al., 2024). While these methods are effective at

explaining and verifying generation, we feel that global index-time attribution is desirable for credit attribution. We discuss this comparison further in Appendix C.

4 Credit Attribution Methods

How should credit for an answer be divided among publishers? There are three natural credit attribution rules: divide credit equally among every publisher associated with a block cited in the answer; assign credit to whomever reported the block earliest, or give credit for marginal contribution over what others could provide. This section shows that all three emerge from a stylized model of news reading as we vary the model parameters. Section 4.5 then shows that these attribution rules can be interpreted as Shapley values of variations of a content generation game.

4.1 The three candidate rules

Fix a query. Let $\mathcal{B}$ be the set of blocks cited in the answer, $n = |\mathcal{B}|$, and for each block $b \in \mathcal{B}$ let $P_b \subseteq \mathcal{P}$ be its *covering set*: the publishers whose articles contain that block. Let $t_{p,b}$ be the time at which publisher p published its article carrying b. One unit of value is to be divided among publishers for each query. An attribution rule φ assigns to each publisher p a share of this value, $\varphi_p \in [0, 1]$. We focus on the following three candidate rules.

Equal. Each block is worth $1/n$ and is split evenly among its publishers:

$$\varphi_p^{\textsc{Equal}} = \frac{1}{n} \sum_{b \in \mathcal{B}} \frac{\mathbb{1}\{p \in P_b\}}{|P_b|}.$$

Provenance. Each block's value goes entirely to whoever published it first, with ties split evenly:

$$\varphi_p^{\textsc{Provenance}} = \frac{1}{n} \sum_{b \in \mathcal{B}} \frac{\mathbb{1}\{p \in \arg\min_{q \in P_b} t_{q,b}\}}{|\arg\min_{q \in P_b} t_{q,b}|}.$$

Completer. Each publisher is, loosely speaking, credited for the unique blocks it contributes. (Ghorbani and Zou, 2019; Jia et al., 2019). Unlike the other two, this rule is typically defined as a Shapley value rather than expressed in closed-form. Consider the cooperative game in which a coalition $S \subseteq \mathcal{P}$ has value $v_{all}(S) = \mathbb{1}\{S \cap P_b \neq \emptyset \text{ for all } b \in \mathcal{B}\}$ if it can supply *every* block of the answer and 0 otherwise. Let σ be a permutation of publishers, $S_{\sigma(p)} = \{p' : \sigma^{-1}(p') < \sigma^{-1}(p)\}$ be the publishers that appear before p in σ,[3] and let

$$\varphi_p^{\textsc{Completer}} = E_\sigma[v_{all}(S_{\sigma(p)} \cup \{p\}]) - v_{all}(S_{\sigma(p)})]$$

[3]Note $S_{\sigma(p)}$ can equivalently be defined as a random subset of publishers excluding p; the definition in terms of permutations is helpful for our characterization of attribution rules in a common framework and also useful for computing the Shapley value.

be the Shapley value of publisher p. An answer that is missing anything is worth nothing, so a publisher earns only in those permutations in which it is the one that completes the set. The rule has no closed form and we evaluate it by sampling permutations.

Remark 1. Because v_{all} is equal to 1 only when *every* block is covered, publishers that supply duplicated blocks only receive credit for such a block in permutations in which they are the first among the duplicate publishers. The residual can be small, as we report in Section 5.2.7 on our empirical data. The rule is therefore close to being a rule that pays for exclusivity alone. Exclusivity is often a knife-edge property of the index rather than of the world. Adding just one outlet carrying an otherwise exclusive block removes an existing publisher's exclusivity; failing to merge two wordings of the same fact into a single block can create exclusivity. Under EQUAL an error of that kind moves a credit allocation from $1/|P_b|$ to $1/(|P_b| \pm 1)$ and washes out across blocks; under COMPLETER it is the difference between a large share and nothing. This also makes the rule susceptible to manipulation, since the way to become indispensable is to publish what nobody else will match, which rewards being hard to cluster rather than being first or being right. For these reasons, we view the completer rule as inherently less robust than the other options.

4.2 A Unifying Framework: The Reading Model

Consider a stylized reader—think of her as the buyer's proxy—who is handed the grounded sources for a query and works through them in some order π. She begins knowing nothing. Each time she finishes a source she registers whichever of its blocks are new to her; blocks she has already seen elsewhere add nothing. Her satisfaction with what she has learned so far is $u(x)$, where $x \in [0, 1]$ is the fraction of the answer's blocks she has covered, with $u(0) = 0$ and $u(1) = 1$. We take

$$u(x) \;=\; x^{\alpha}, \qquad \alpha \geq 1, \tag{1}$$

a one-parameter family that spans the range of views a buyer might hold about an incomplete answer. At $\alpha = 1$ utility is linear and every additional fact is worth the same. As α rises the early facts count for less and the last ones for more, and in the limit $\alpha \to \infty$ we have $u(x) \to 0$ for every $x < 1$ while $u(1) = 1$: the buyer has no value for the answer before consuming all its blocks. The exponent therefore can be interpreted as controlling *how much the buyer values complete answers* and it is a parameter of our model.

Each of the aforementioned attribution rules can be interpreted as assigning a publisher credit for the increment in the reader's satisfaction that it caused. Writing $c_p(\pi)$ for the number of blocks already covered when publisher p is reached under order π,[4] and $a_p(\pi)$ for the number

[4]One might read the sources in a random order over *articles* rather than over *publishers*, which is not the same thing when a publisher contributes several articles. In our data the two differ by 0.14% of total mass, so we use the publisher-level version throughout and report the article-level version in Appendix B.

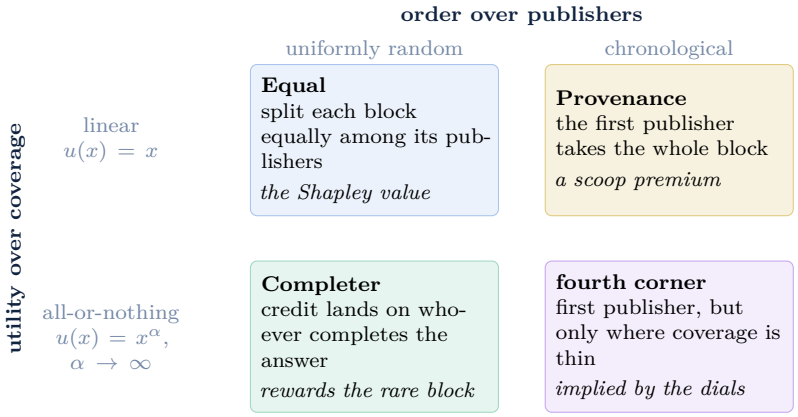


Figure 4: Credit Allocation Rules. Methods for dividing credit among publishers who contributed a duplicated block, expressed as a function of reading order and utility curvature. Three of the corners are the rules described in Section 4.1; the fourth is implied by the framework.

of blocks p is the first to supply,

$$\varphi_p \;=\; \mathbb{E}_\pi\Big[\, u\Big(\frac{c_p(\pi) + a_p(\pi)}{n}\Big) - u\Big(\frac{c_p(\pi)}{n}\Big)\Big]\,. \tag{2}$$

Equation (2) describes the mechanism. It has exactly two free parameters (or "dials"): the distribution over reading orders π, and the exponent α. These two parameters determine the credit attribution rule. Figure 4 illustrates the construction.

4.3 Varying the order parameter: distribution over π

First fix the exponent $\alpha = 1$ so that utility is linear in the fraction of blocks read. If the reading order π is chosen uniformly randomly over all permutations, a block is credited to whichever of its publishers the reader happens to reach first, and by symmetry that is each of them with probability $1/|P_b|$. In expectation every publisher in P_b receives $1/(n|P_b|)$—which is the EQUAL rule exactly.

Proposition 1 (Random order gives the equal split)**.** *With u linear and π uniform over orderings of the covering publishers, $\varphi_p = \varphi_p^{\textsc{Equal}}$ for every p.*

PROVENANCE is the *same* mechanism with the order dial turned to history: π is chosen to order articles by the time at which they are published. Recall that under linear u the credit for block b goes entirely to whichever member of P_b the reader reaches first. When π is chronological that member is the earliest publisher of b by construction, so (2) pays $1/n$ to the originator of each block and nothing to its later publishers, yielding the PROVENANCE rule.

Proposition 2 (Chronological order gives the provenance rule)**.** *With u linear and π fixed to be the chronological order of publication, $\varphi_p = \varphi_p^{\textsc{Provenance}}$ for every p.*

Thus the two rules differ only in their distribution over π. One can interpret a uniformly random order π as a veil of ignorance: it says the market should not reward a publisher for reporting on a fact first, only for being able able to supply the fact at all. Chronological order says the opposite: being first is what carries value. Intermediate rules, taking distributions over π that favor provence without insisting on it for example, provide an interpolation between these extremes. The choice of distribution is a normative choice about whether original reporting or availability is what the marketplace exists to fund. In our empirical study, we find that choosing one extreme or the other can shift aggregate revenue share of a publisher by several percentage points. See Section 5.2.

4.4 Varying the utility model: parameter α

Suppose that the order π is chosen uniformly at random and consider varying the parameter α. The exponent α of (1) runs from $\alpha = 1$, where every additional fact is worth the same to the buyer, to the limit in which u tends to $\mathbb{1}\{x = 1\}$ and only a *complete* answer is worth anything. At that limit a publisher's marginal contribution is one precisely when it is the arrival that completes the set, and zero otherwise. Coupling σ from the definition of $\varphi_p^{\textsc{Completer}}$ and π and noting that $v_{all}(S) = u(\{b|P_b \cap S \neq \emptyset\})$ thus implies the following proposition.

Proposition 3 (All-or-nothing utility gives the completer rule)**.** *With π uniform over orderings, $\varphi_p \to \varphi_p^{\textsc{Completer}}$ for every p as $\alpha \to \infty$.*

Equal and Completer are therefore the two ends of one continuum rather than alternatives, and intermediate values of α trace the path between them. Figure 9, in Section 5.2, follows that path through our data. The exponent is the buyer's answer to a single question—*how much is a partial answer worth?*—and the two rules represent two extreme points.

The space of parameters suggests four corner mechanisms. We have already covered Equal, Provenance and Completer. The fourth rule, chronological order with an all-or-nothing buyer, pays the publisher whose arrival completes the answer, where arrivals are ordered by publication time. It therefore rewards whoever was slowest among those the answer could not do without. We report its payouts on our empirical data in Section 5.2.

4.5 Connection to Shapley value

In a cooperative game, a characteristic function describes the value of any coalition, or subset, of players. The Shapley value of a cooperative game assigns credit to individuals by considering their marginal contribution to the characteristic function over a uniformly-random subset of players. It is the unique rule that satisfies *efficiency* (the sum of players' credits equals the total value of the grand coalition), *symmetry* (if two players are interchangeable in all coalitions, they receive the same credit), *additivity* (for games described by characteristic functions v and w, the Shapley values of game that is a linear combination of v and w is the linear combination of the

Shapley values of v and w), and the *null player property* (if a player contributes nothing to any coalition, they get no credit). As such, it is a popular solution concept in practice, as discussed in Section 2.

In this section, we show the rules discussed previously are in fact Shapley values of appropriately-defined characteristic functions.[5] Doing so lets us understand the normative consequences of one choice over another by examining the characteristic function. Recall that COMPLETER was defined as a Shapley value in Section 4.1 and Proposition 3 shows the reading model recovers it. The other two rules were derived from the reading model directly. We now note that they can likewise be interpreted as Shapley values under certain game variations.

Write $v_{\text{lin}}(S) = \frac{1}{n}\,|\{b : S \cap P_b \neq \emptyset\}|$ for the share of the answer's blocks a coalition S can cover. Where v_{all} asks whether S can supply everything, v_{lin} asks how much of it S can supply.

Proposition 4 (The equal split is a Shapley value)**.** *The EQUAL rule is the Shapley value of* v_{lin}*.*

Proof sketch. Write $v_{\text{lin}} = \frac{1}{n}\sum_b v_b$ where $v_b(S) = \mathbb{1}\{S \cap P_b \neq \emptyset\}$ is a simple game in which only the members of P_b matter. Any publisher outside P_b is a null player in v_b and receives nothing; the members of P_b are symmetric in v_b; efficiency then forces each to receive $1/|P_b|$. The result then follows by additivity of the Shapley value. The full proof is in Appendix A. □

PROVENANCE is likewise a Shapley value, of a third game which defines the value of a coalition according to what share of the answer the group *broke*. Writing p_b for the publisher that reported block b first,

$$v_{\text{prov}}(S) \;=\; \frac{1}{n}\,\big|\{b \in \mathcal{B} :\, p_b \in S\}\big|, \tag{3}$$

so a group is credited with a block when one of its members was the first to report it, and carrying a block somebody else broke is worth nothing.

The consequence is that a publisher which only ever reprints contributes nothing to *any* coalition, rather than merely to some of them. Suppose three outlets carry a block and outlet A filed first. Under v_{lin} the singleton $\{B\}$ is worth $1/n$: with A absent, B can still report the block to the reader. Under v_{prov} the singleton $\{B\}$ is worth zero, because the block was not B's to break, and adding B to any group leaves that group's value unchanged. So B is a null player and A is the only publisher whose presence or absence moves the block at all. The Shapley value pays a null player nothing and hands the block's whole value to the publisher that alone determines it, which is the provenance rule.

Proposition 5 (The provenance rule is a Shapley value)**.** *The PROVENANCE rule is the Shapley value of* v_{prov}*.*

We can interpret $v_{\textbf{prov}}$ as saying a reprint is worth nothing to the buyer even when the original is unavailable. Taken literally that is false, and it is exactly the same commitment the chronological

[5] Note this is non-trivial because Shapley value requires the marginal contribution to be computed over a uniformly-random subset or, equivalently, a uniformly-random permutation.

order makes. There the later publisher earns nothing because it is always reached second, here because it never counted. This is two ways of saying the same thing: either the buyer values coverage and reads in the order the news broke, or the buyer reads in no particular order but values only first publication. Timestamps enter the mechanism at one point, and whether one puts them into the reading order or into the game is immaterial; the credit allocations are the same either way.

5 Empirical Study

5.1 The setting and data

Our data come from a deployed prototype of a block-level marketplace. Our domain is news articles covering the Iran/US war, available online on June 28, 2026. A ceasefire had been in effect at the time, but Iran attacked container ships in the strait of Hormuz during the weekend and the US retaliated. We selected this episode because it is fast-moving, heavily covered, and generated a large volume of near-simultaneous reporting by outlets of very different types, so both original reporting and copying are present in quantity.

To construct our corpus, we sampled 1,994 user queries posed to a major search engine about this news scenario over a 1-hour period on June 28, 2026. Our corpus consists of all articles, records, or URLs returned from grounding in response to these queries. Of the 1,994 user queries, 1,974 produced an answer citing at least one block; 23,583 grounding records over 7,553 distinct URLs; and 1,375 publishers, of whom 137 (10.0%) are classified as premium. Articles decompose into 377,238 block instances which cluster into 64,419 distinct blocks. The mean answer cites 10.66 blocks. Figure 5 illustrates the resulting distribution of cluster sizes across news and commentary blocks for all extracted claims versus those cited in generated answers, for the set of articles crawled.

5.1.1 Premium classification

Throughout we report the share of value accruing to *premium* publishers. This is a designation intended to capture established outlets with substantial original newsgathering operations, in contrast to aggregators, syndicators, and low-cost republishers.

The premium coding is a hand-curated tier assigned to each publisher domain, with three levels. 54 domains are coded `premium-us` and 83 `premium-intl`; the remaining 1,238 are `standard`. The two premium levels are pooled throughout and reported as *premium*, giving the 137 publishers (10.0%) that the results tables refer to. The list is dominated by national newspapers, wire services and broadcasters. 55 of the premium domains are American, 15 French, 13 Indian and 9 British. Importantly, the tier was assigned before any attribution rule was computed, so it was not fitted to the results.

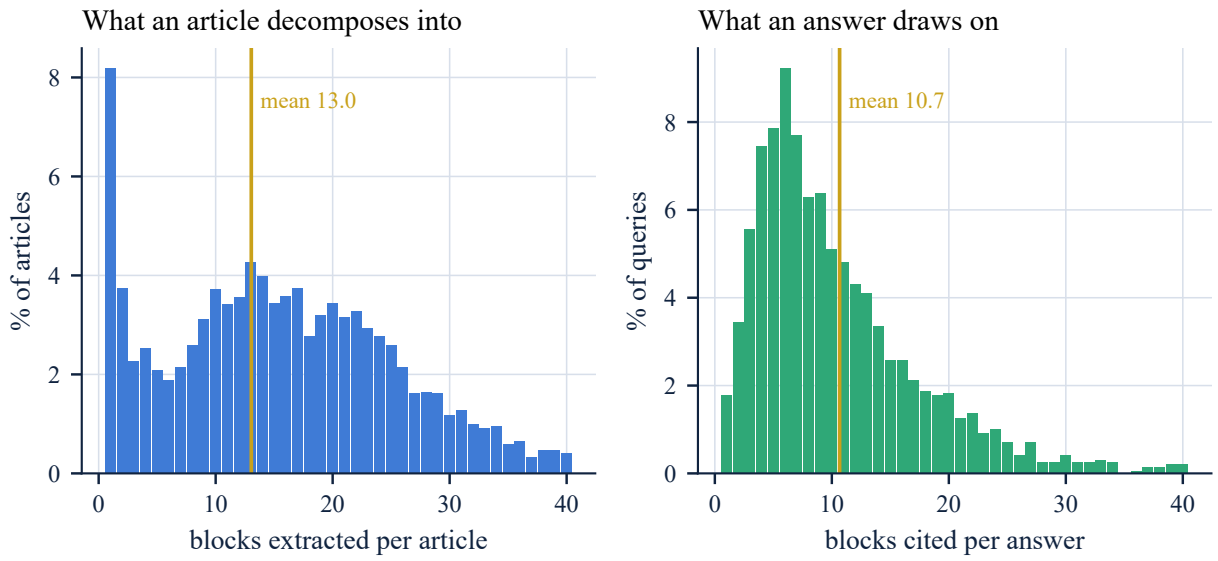


Figure 5: Anatomy of the index. Left: blocks extracted per article. Right: blocks cited per answer. Both truncated at 40 for legibility.

The premium classification is not a normative judgment about quality. It is a proxy for the cost structure the marketplace is meant to affect: those publishers denoted "premium" are those that presumably pay reporters to generate content. As a robustness check, we report in Appendix B how the results change when the two premium levels are separated and when the premium tier is replaced by the fully mechanical *type* coding, which identifies wire services, national news, and broadcast outlets.

Two caveats about the set of publishers used in our data. First, publishers are identified by domain, so a small number of outlets appear more than once under country editions: 16 common names are carried by two or more domains. This splits their credit and therefore mildly *understates* concentration under every credit attribution rule. Second, 1,126 of the 1,375 domains fall in the residual *other* type under our type coding, which is a reminder that the long tail of the crawl over news articles is primarily covered by small publishers that are not indexed as major news organizations.

5.1.2 Publication timestamps

The PROVENANCE rule requires knowing the order of publication. Of the 7,553 grounded URLs, 87.2% carry some publication time. But a large fraction carry only a *date*, which the crawler records as midnight. 3,043 URLs have a recorded time but it sits at exactly 00:00:00, suggesting a similar truncation, and a further 966 have no time at all. The share carrying an exact, minute-resolution time is 33.3% for premium publishers and 58.4% for the rest.

We emphasize that this the asymmetry runs *against* premium outlets in the PROVENANCE rule, since they are the group whose exact filing times are least often recorded. Furthermore,

Table 1: Premium share of value, global index, by timestamp convention.

Timestamp convention	EQUAL	PROVENANCE
(no timestamps used)	41.87%	—
Times as stored	—	49.55%
Date-only imputed to noon	—	49.11%
Truncated to calendar day	—	46.79%
Exactly-timed publishers only	—	46.33%

1,974 queries. Ties are split equally under every convention. EQUAL does not use timestamps and so is invariant to the convention.

because PROVENANCE is decided by an ordering, and because a third of the relevant timestamps are imputed, our results for PROVENANCE are potentially sensitive to the imputation convention. We therefore consider and compare four conventions in our reported results: times as stored; date-only records imputed to noon rather than midnight; all times truncated to the calendar day so that same-day publications tie; and exact times only, with blocks lacking any exactly-timed publisher split equally.

5.2 Empirical Results

5.2.1 Prevalence of duplication among blocks

Among cited blocks, evaluated against the global index, the mean number of carrying publishers is 22.4 and only 27.2% have a single publisher. Duplication is also unevenly distributed across the two block kinds. News blocks are singletons 10.5% of the time, commentary blocks 49.3% of the time, meaning that facts get copied more often than the commentary written about them. Figure 7 reports the full distributions.

Due to the prevalence of duplication, we find that deduplicating facts substantially reduced context length. Across all 1,994 queries in our evaluation, generating answers from block trees consumed 16,000,243 prompt tokens compared to 35,067,592 for the raw-text baseline—a 54.4% reduction in prompt overhead. Including generated completions, total token consumption fell by 52.3% (Figure 6).

5.2.2 Provenance favors premium publishers

Table 1 reports the premium share of value under the EQUAL and PROVENANCE rules,

computed against the global index, under each of the four timestamp conventions of Section 5.1.2.

Rewarding first publication moves between five and eight percentage points of total value toward premium publishers. The sign and magnitude are roughly constant across timestamp conventions. The mechanism is priority: premium outlets are disproportionately the ones who

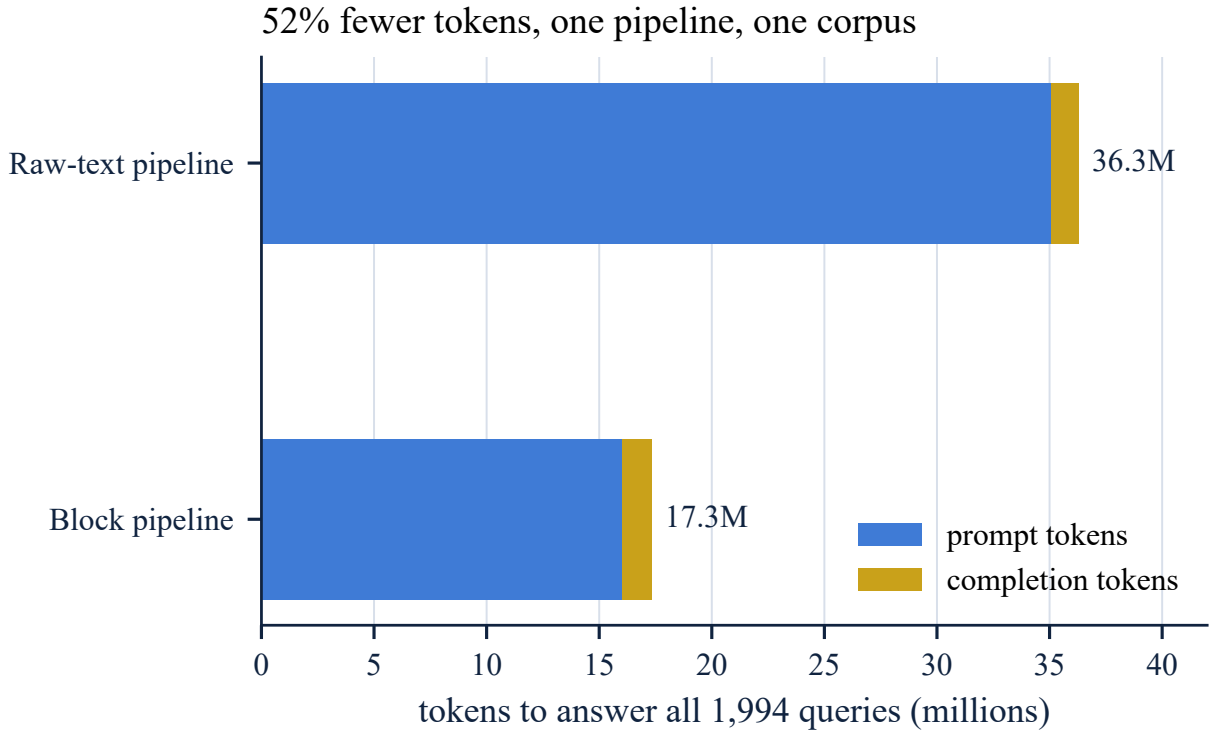


Figure 6: The answering cost (C_2). Prompt and completion tokens per query, comparing deduplicated block trees against raw article text across 1,994 queries.

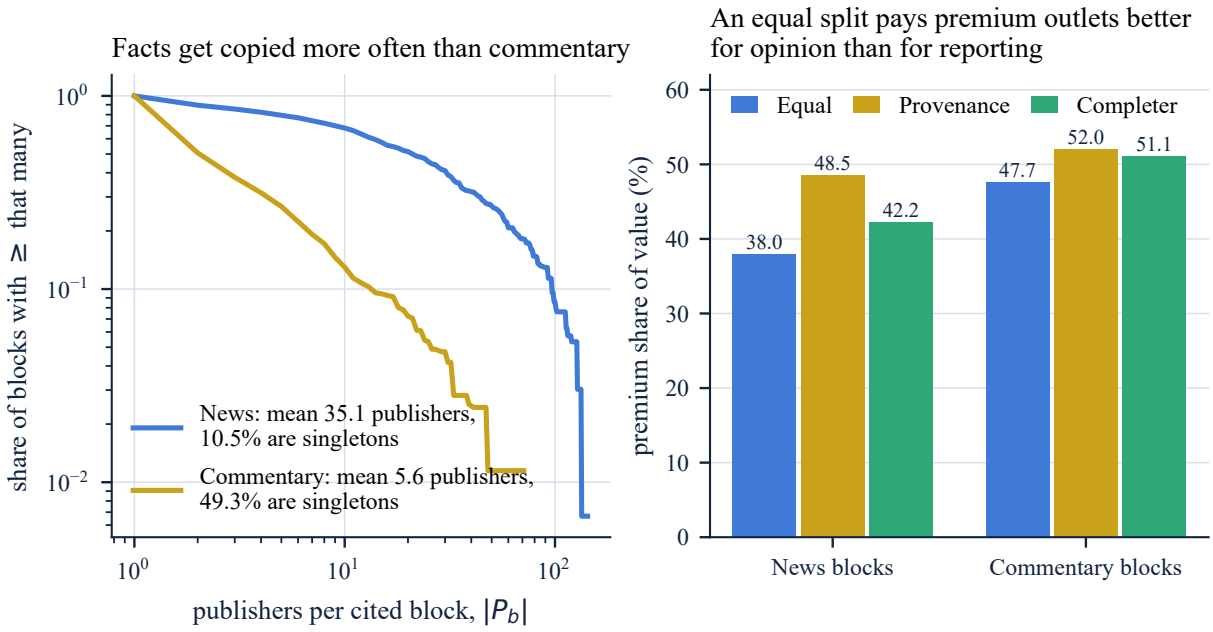


Figure 7: The duplication gap. Left: survival function of publishers per cited block in the global index, news against commentary. Right: the same gap in payment terms—an equal split pays premium outlets nearly ten points more for opinion than for reporting.

Table 2: The global and local index compared.

	Local index	Global index
Mean publishers per cited block	1.23	22.4
Single-publisher blocks	86.2%	27.2%
EQUAL, premium share	53.75%	41.87%
PROVENANCE, premium share	53.4–54.1%	46.3–49.6%

PROVENANCE ranges are across the four timestamp conventions of Table 1.

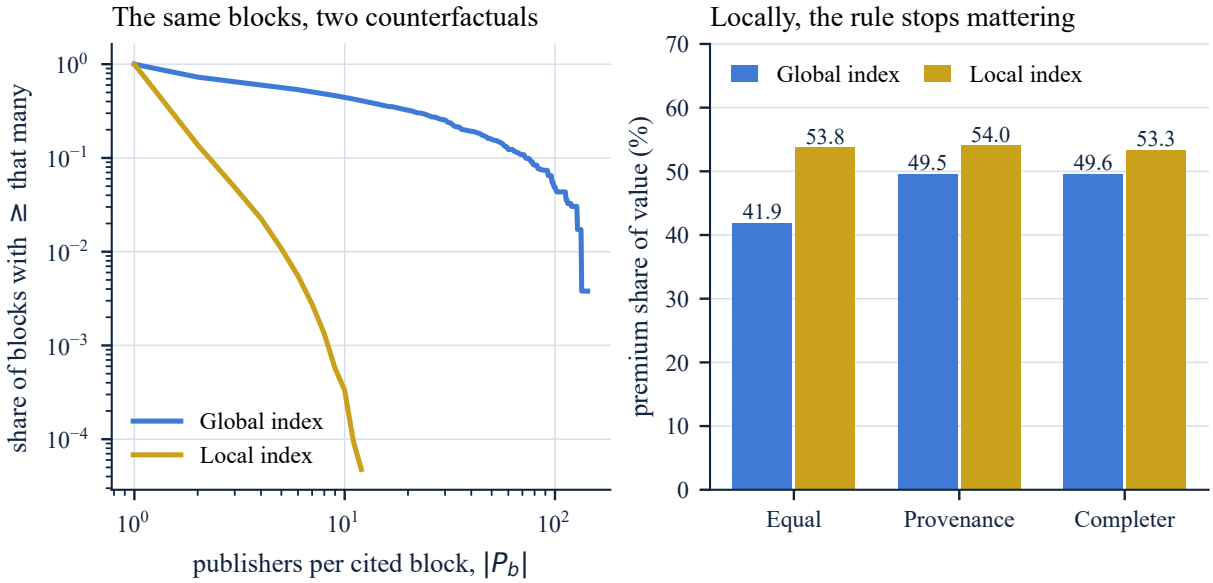


Figure 8: Global against local index. Left: coverage distributions under the two scopes. Right: the premium share under each rule. Locally, the rules collapse onto each other—there is nothing left for a payment rule to disagree about.

report a fact first and are then carried by others, so a rule that prices availability and a rule that prices priority reach materially different answers.

5.2.3 The impact of global indexing

Table 2 recomputes the same quantities as Table 1 but using the local index instead of the global index.

Inside the local index, 86.2% of cited blocks have exactly one publisher and mean coverage is 1.23. Since duplication is rare, all rules arrive at approximately the same outcome. As a result, every rule and every timestamp convention lands within a single percentage point of every other. Figure 8 shows the two coverage distributions side by side.

We regard this as an important empirical result. At the scope of the local index, **no attribution policy is expressible at all**. Since an inference-time method can only ever observe the retrieved set, different attribution strategies evaluated at inference time against the local

Table 3: The utility dial: premium share of value, global index, random order.

Buyer's utility	Premium share	Publisher-level distance from EQUAL (TV)
$\alpha = 1$ (EQUAL, closed form)	41.87%	—
$\alpha = 1$ (sampled)	41.88%	0.57 pp
$\alpha = 2$	45.11%	
$\alpha = 4$	47.30%	
$\alpha = 8$	48.64%	
$\alpha \to \infty$ (COMPLETER)	49.56%	19.6 pp

Distances are total variation between normalized publisher payout vectors, computed over all publishers, so a distance can be large even where the premium aggregate agrees: individual errors cancel within the group. The second row is the Monte Carlo estimate of a quantity the first row computes exactly; the gap between them is sampling noise and bounds the precision of the rest of the column. Appendix B.4 reports the full convergence, which follows the textbook $m^{-1/2}$ rate.

index will result in very similar payouts to publishers.

5.2.4 Comparing the Equal and Completer Rules

Next we compare the EQUAL and COMPLETER rules. Payments are the expected marginal contributions as described in Section 4, estimated by sampling 200 publisher orderings per query. The results are reported in Table 3 and Figure 9.

Premium share rises monotonically in α, from 41.87% to 49.56%. Most of the movement happens early: going from $\alpha = 1$ to $\alpha = 2$ buys more than half the total change. Curvature in the buyer's utility is a preference for completeness, completeness is delivered by whoever holds the blocks nobody else has, and premium publishers hold disproportionately many of those.

It is worth noting that both COMPLETER and PROVENANCE increase payments to premium publishers relative to EQUAL, and by nearly the same amount in aggregate. COMPLETER pays premium publishers 49.56% and PROVENANCE pays them 49.55%, a difference of less than a tenth of a point. These are economically distinct rules—one rewards being irreplaceable, the other rewards being early—and they agree on aggregate share to premium publishers. They do not, however, agree on how payoffs are distributed among individual publishers. The total variation between their publisher-level payout vectors is 33.7 percentage points, larger than the distance from either of them to EQUAL.

5.2.5 The fourth corner, and the two random orders

Two further outcomes are worth discussing. Combining the chronological order with the all-or-nothing utility gives the fourth corner—pay whoever closes the last gap, in historical order—which pays premium publishers 47.57%, inside the range spanned by the three named rules.

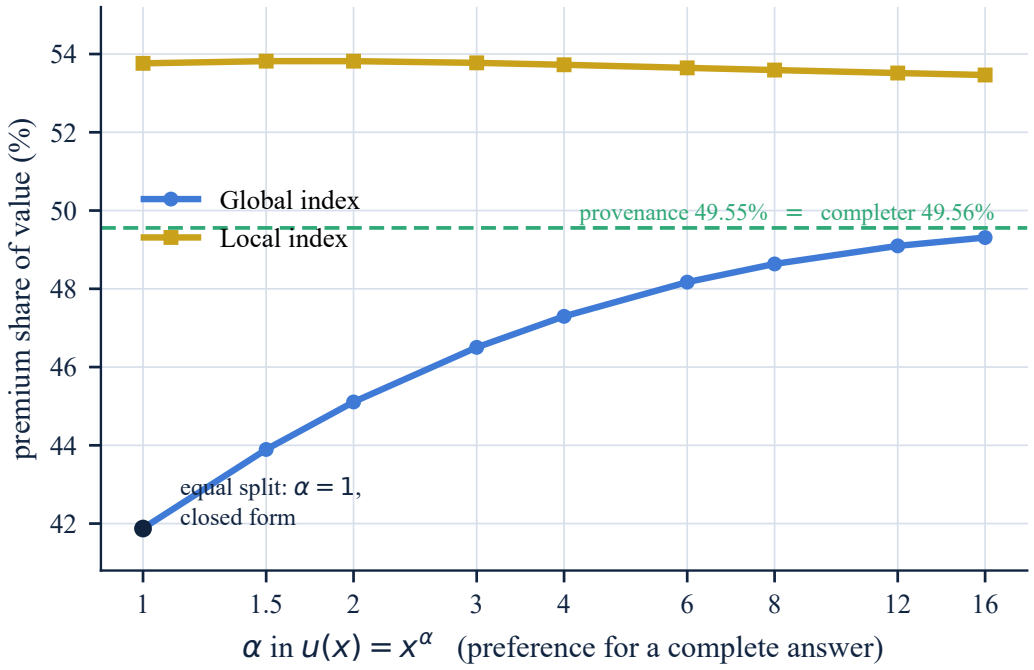


Figure 9: The utility dial. Premium share of value against α in $u(x) = x^{\alpha}$, random order, under both index scopes. Provenance and the completer land within a hundredth of a point of one another and are drawn as a single reference line—from opposite economic directions.

Separately, Section 4.3 distinguished two ways of randomizing: over *articles* and over *publishers*. We define the EQUAL rule as randomizing over publishers. Empirically the distinction is immaterial in the local index, where the two differ by 0.07 percentage points, and not immaterial in the global index, where they differ by 5.57 points and the premium share moves from 41.87% to 45.26%. Randomizing over articles gives extra weight to publishers who filed many stories carrying the same fact, which is a form of volume subsidy. This suggests that article-level randomization could implicitly reward output and repetition rather than content.

5.2.6 The news and commentary split

The premium share differs sharply by block kind, and the difference tracks the duplication gap of Section 5.2.1.

A cited news block is carried by 35.1 publishers on average; a cited commentary block by 5.6. Under EQUAL, premium publishers receive 37.95% of the value of news blocks but 47.66% of the value of commentary—an equal split pays them substantially better for opinion than for reporting, because opinion is not copied and so is not diluted. Moving to PROVENANCE raises the news figure to 48.51% while raising the commentary figure only to 52.01%, which closes most of the gap.

This has implications for the impact of rule choice on the relative attribution of credit across content types. **An equal split does not merely fail to reward original reporting; it**

Table 4: Premium share of value by block kind, global index.

	News	Commentary
Cited blocks	11,991	9,059
Mean publishers per block	35.1	5.6
Single-publisher blocks	10.5%	49.3%
EQUAL, premium share	37.95%	47.66%
PROVENANCE, premium share	48.51%	52.01%
COMPLETER, premium share	42.24%	51.09%

Table 5: Concentration of the pool across publishers, global index.

	EQUAL	PROVENANCE	COMPLETER
Top 10 publishers' share	23.2%	34.6%	33.4%
Top 50 publishers' share	50.1%	68.5%	63.9%
Gini coefficient	0.748	0.885	0.844
Publishers paid nothing	0.0%	34.6%	12.3%

Base is the 1,103 publishers who cover at least one cited block; publishers earning nothing under a rule remain in the denominator, so the columns are comparable. Under the local index the three rules are near-indistinguishable, with EQUAL at a Gini of 0.840 and PROVENANCE at 0.846.

downweights reporting relative to commentary, because the property that makes reporting valuable—that others find it worth copying—is the same property that dilutes its payment.

5.2.7 Concentration

We next analyze how widely payments are dispersed among publishers. This is reported in Table 5. The top fifty publishers take 50.1% of the pool under EQUAL and 68.5% under PROVENANCE. We use the Gini coefficient to measure concentration of payments, and find that it rises from 0.748 to 0.885. Rewarding originality is therefore not a transfer from the long tail to premium outlets in the aggregate sense only, but it is also a substantial concentration of the pool. The aggregate premium share understates this concentration because the movement is toward a small number of large originators rather than toward the premium tier as a class.

Under EQUAL, every publisher who covers a cited block is paid something. Under PROVENANCE, 34.6% of them are paid nothing at all, because being second is worth exactly zero. This is not a distributional nuance; it is a participation question. A marketplace in which a third of contributing publishers receive no payment is one that a third of publishers have no reason to join, and the value of the index to the buyer depends on their joining.

The completer rule concentrates for a different reason, as discussed in Section 1. Because v_{all} is one only when every cited block is covered, a publisher that uniquely supplies some cited block

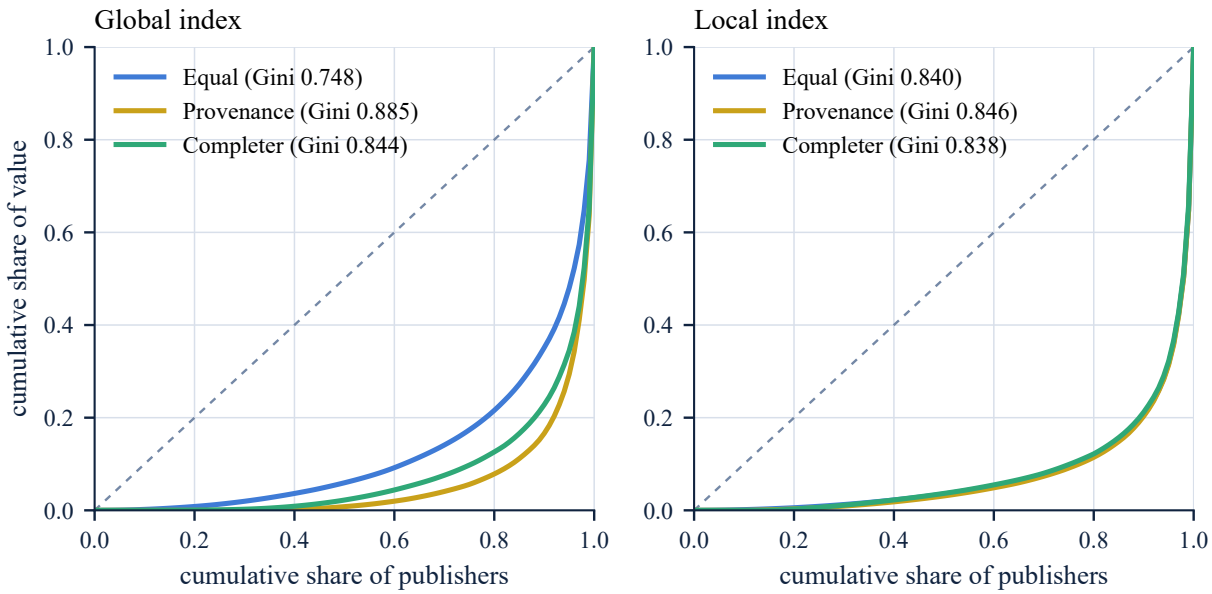


Figure 10: Concentration. Lorenz curves of value across publishers under each rule, on a fixed base of contributing publishers. Under the local index the three curves are indistinguishable.

is a veto player, and veto players collect 66.3% of the pool under the global index. Under the local index, where 86.2% of blocks have a single publisher, they collect 97.1%. The completer rule is thus close to being a rule that pays only for exclusivity, and the local index turns it into one almost exactly. This is our main reason for treating it as an extremal example within our framework rather than as a candidate policy.

Figure 10 reports the full distribution of value across publishers under each rule, with the premium tier overlaid, and Figure 11 reports the veto-player share alongside the top-k curves.

5.2.8 Publisher-level results

So far, all results have been reported in aggregate. Table 6 reports anonymized publisher-level results, since the rules are easier to judge against individual payout patterns than against a Gini coefficient.

The top of the equal split is held by outlets that publish prolifically. Premium publisher/broadcast #1, premium publisher/broadcast #2, general-interest platform #1, premium publisher/broadcast #3 and standard publisher #1 lead it. That is the rule working as designed: it pays for having the fact, and these are the outlets that have the most facts.

Premium publisher/broadcast #9 is the cleanest illustration of the distinction between Equal and Provenance. It is 29th under Equal with 0.62% of the pool and 3rd under Provenance with 4.70%, a factor of 7.6. The mechanism is the dilution of Section 5.2.1: facts that an outlet originates may be carried by many other outlets, so an equal split divides each of them twenty-odd ways. General-interest platform #1 moves the other way, from 2.95% to 1.49% and from third to

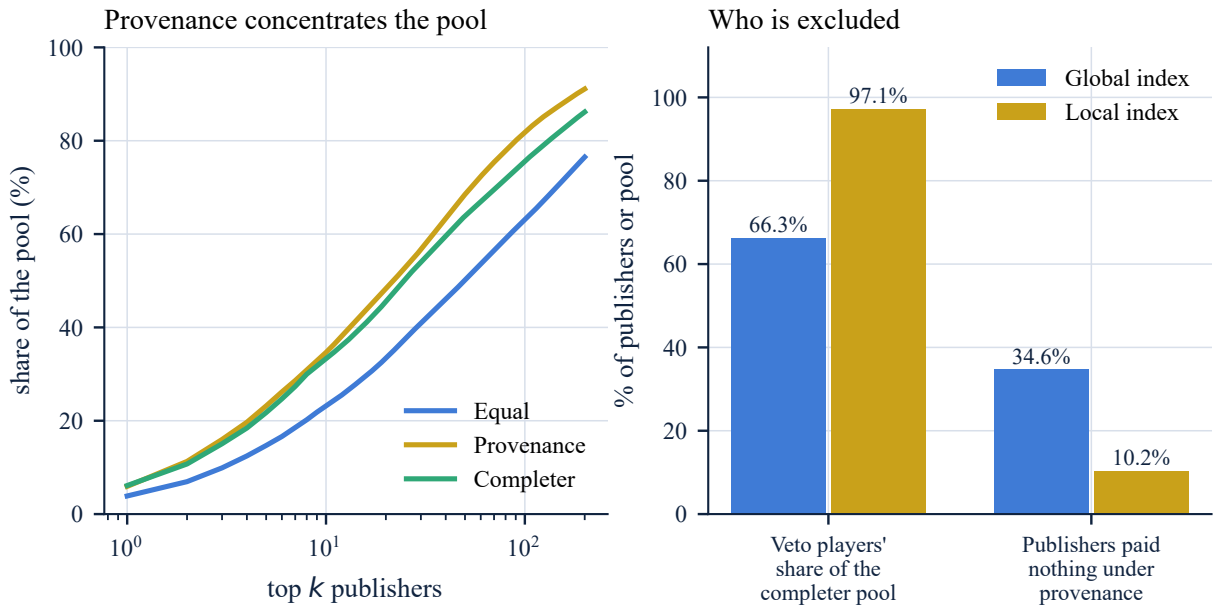


Figure 11: Who gets paid. Left: top-k share of the pool under each rule, global index. Right: the share of the completer pool collected by veto players, and the share of contributing publishers paid nothing at all under provenance.

fourteenth. Premium publisher/broadcasts #10 and #11 rise under Completer without rising under Provenance—#10 from 1.21% to 2.94%—which is the signature of holding material nobody else carried rather than of having been early.

Another row from the table is worth highlighting. Specialized information site #1 is 95th under Equal with 0.19% of the pool and 4th under Provenance with 3.70%, which puts a single-purpose site ahead of almost every newsroom in the corpus. We do not believe this specialized information site was the first to report these facts. We believe it carries filing times that sort early, as discussed in Section 5.1.2. This illustrates the potential fragility of a first-mover rule: it is only as good as the timestamps it is decided by.

5.3 Empirical Study Limitations

Our evidence comes from a single news episode served by a single retrieval and generation stack. The Iran/US coverage was chosen because heavy simultaneous reporting makes duplication measurable, but that same feature may make our covering sets unusually large. We would expect the qualitative results—that duplication is common and that the index scope dominates the rule—to generalize.

Publication timestamps are substantially imputed. As Section 5.1.2 sets out, only 33.3% of premium and 58.4% of non-premium grounded URLs carry an exact time. As a robustness check, we report four different conventions and verify that our results are largely stable across them. That said, we note that this robustness may not generalize to other datasets, so further study on

Table 6: The ten largest recipients under EQUAL, and what the same anonymized publisher groups are paid under the other two rules.

Publisher	EQUAL	PROVENANCE	COMPLETER
Premium publisher/broadcast #1	3.88 (1)	5.35 (2)	6.12 (1)
Premium publisher/broadcast #2	3.07 (2)	3.50 (5)	4.62 (2)
General-interest platform #1	2.95 (3)	1.49 (14)	4.33 (3)
Premium publisher/broadcast #3	2.56 (4)	5.93 (1)	3.37 (4)
Standard publisher #1	2.24 (5)	1.72 (12)	3.31 (5)
Premium publisher/broadcast #4	1.92 (6)	3.00 (6)	2.66 (7)
Premium publisher/broadcast #5	1.88 (7)	2.43 (7)	2.63 (8)
Premium publisher/broadcast #6	1.70 (8)	1.92 (9)	1.74 (9)
Premium publisher/broadcast #7	1.63 (9)	1.18 (17)	1.17 (16)
Premium publisher/broadcast #8	1.32 (10)	2.23 (8)	1.29 (14)
Outside this ten, but inside the top ten under one of the other rules			
Premium publisher/broadcast #9	0.62 (29)	4.70 (3)	0.66 (30)
Specialized information site #1	0.19 (95)	3.70 (4)	0.16 (99)
Reference platform #1	0.92 (18)	1.88 (10)	1.03 (22)
Premium publisher/broadcast #10	1.21 (11)	1.31 (16)	2.94 (6)
Premium publisher/broadcast #11	1.10 (14)	1.37 (15)	1.64 (10)

Percent of the whole pool, global index, with PROVENANCE under times as stored. Rank among the 1,103 contributing publishers in parentheses. Labels are anonymized for this version of the paper.

the impact of imputation methods is warranted.

A deeper point about ordering is that we observe *who filed first*, which is not the same as *who found out first*. Two outlets may work the same story independently and file an hour apart; one may lift the other's reporting wholesale and file a minute later. The index records both as one cluster with two publishers and a filing order, and nothing in our data distinguishes independent discovery from copying. PROVENANCE therefore prices *priority*, not *originality*, and reads as a rule about original reporting only under the assumption that the two coincide. That assumption is plausible in a fast-moving news episode, where being first usually does mean having done the work, but it is an assumption.

The index is also bounded by the crawl, and our estimates condition on retrieval. The global index is global relative to what was crawled, not relative to everything published. The direction of this bias is signable: a larger corpus can only increase the mean number of publishers per block, which can only widen the gap between rules that price availability and rules that price originality. Our estimates of that gap are therefore best read as lower bounds. Separately, all our rules are computed over blocks that were actually *cited*, so they still condition on the retriever's choices. The framework does not require this—covering sets are defined over the whole index—but our measurements inherit it.

A further constraint is extraction and clustering error. The block index is produced by a

language model and inherits its mistakes, both in decomposing articles and in deciding that two blocks express the same content. Under-clustering inflates the singleton share and biases toward rules that reward indispensability; over-clustering does the reverse. Appendix B perturbs the index directly along all three error modes—misfiled kinds, clusters wrongly split, clusters wrongly merged—at rates up to 20%. We view this as a reasonable upper bound on plausible error rate for this pipeline. Our main results are orobust to these perturbations: the PROVENANCE–EQUAL spread stays between +7.62 and +8.12 percentage points against a baseline of 7.68.

Finally, certain potentially relevant parameters are not observable in our dataset. We take answer quality as given and make no claim that block-based generation produces better answers, only cheaper ones. We do not model publishers' entry or investment decisions, so any discussion of publisher incentives is informal rather than an equilibrium analysis. We also do not observe prices: we allocate shares of a normalized pool, leaving the level of the pool—which is what publishers actually care about—entirely outside the paper.

6 Conclusion

The click is disappearing as a way to compensate publishers for their content, and something has to replace it. While there is broad agreement that publishers should be paid for the content ingested and served by AI to answer queries, the norms and mechanisms by which payments are dispersed have not yet been settled.

We have argued three things. First, a variety of natural credit attribution rules—equal split, first-mover, and completer–emerge as Shapley values in a general framework of news readership games.

Second, we have argued that the distinction between these rules is meaningful in empirical data. We find that attributing credit based on a first-mover rule rather than equal split shifts five to eight percentage points of value toward premium publishers, robustly across four ways of handling incomplete timestamps. This suggests a non-trivial impact on usage payments, though arguably not a complete transformation of investment incentives.

Third, we explored the impact of attributing credit with respect to a *global index* of everything published, at index time, versus evaluating credit against the *local index* of sources retrieved for one query at inference time. In the global index a cited news fact has a mean of 22.4 publishers and the rules separate. In the local index 86.2% of cited facts have a single publisher and every rule results in essentially identical payments. We conclude that a global index is critical for designing allocation policies that meaningfully capture an intended metric of publisher contribution. Attribution computed when a query arrives can only ever see the local index, so it can never price the counterfactual the market cares about—what this publisher is worth given everything else that exists. Attribution computed when a document is indexed can, with the architectural advantage that this determination does not impose a significant computational cost

per query.

Finally, a note on what a marketplace settlement layer requires that a metering scheme does not. A global indexing system can be observed and verified: Publishers can be shown their blocks, who else carried them, and all timestamps. This observability enables contracts written on these verifiable states. The auditability of an index-based measurement system therefore enables contractable payments to publishers along with the feedback needed to help publishers invest in content that is valued. We argue that this is an important step toward developing marketplaces and mechanisms that maintain the long-term health and stability of the content ecosystem in the age of AI-generated answers.

References

Luis Aguiar and Joel Waldfogel. Platforms, power, and promotion: Evidence from Spotify playlists. *Journal of Industrial Economics*, 69(3):653–691, 2021. doi: 10.1111/joie.12263.

Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, 2020. doi: 10.1126/sciadv.aay3539.

Will Allen and Simon Newton. Introducing pay per crawl: Enabling content owners to charge ai crawlers for access. Cloudflare Blog, July 2025. URL `https://blog.cloudflare.com/introducing-pay-per-crawl/`. Accessed 2026-08-23.

Charles Angelucci, Julia Cagé, and Michael Sinkinson. Media competition and news diets. *American Economic Journal: Microeconomics*, 16(2), 2024. doi: 10.1257/mic.20220163.

Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. Should we treat data as labor? moving beyond "free". *AEA Papers and Proceedings*, 108:38–42, 2018. doi: 10.1257/pandp.20181003.

Kenneth J. Arrow. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity*, pages 609–626. Princeton University Press, 1962.

Susan Athey, Markus Mobius, and Jeno Pal. The impact of aggregators on internet news consumption. NBER Working Paper 28746, National Bureau of Economic Research, 2021.

Yannis Bakos and Erik Brynjolfsson. Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 45(12):1613–1630, 1999. doi: 10.1287/mnsc.45.12.1613.

Dirk Bergemann and Alessandro Bonatti. Markets for information: An introduction. *Annual Review of Economics*, 11:85–107, 2019. doi: 10.1146/annurev-economics-080315-015439.

Dirk Bergemann, Alessandro Bonatti, and Tan Gan. The economics of social data. *RAND Journal of Economics*, 53(2):263–296, 2022. doi: 10.1111/1756-2171.12407.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Business Wire. ProRata Invents Generative AI Attribution Technology to Compensate and Credit Content Owners While Facilitating Fairness and Fact. Corporate Press Release, August 2024. URL `https://businesswire.com`.

Julia Cagé. Media competition, information provision and political participation. *Journal of Public Economics*, 185:104077, 2020. doi: 10.1016/j.jpubeco.2019.104077.

Julia Cagé, Nicolas Hervé, and Marie-Luce Viaud. The production of information in an online world. *Review of Economic Studies*, 87(5):2126–2164, 2020. doi: 10.1093/restud/rdz061.

Joan Calzada and Ricard Gil. What do news aggregators do? evidence from Google News in Spain and Germany. *Marketing Science*, 39(1):134–167, 2020. doi: 10.1287/mksc.2019.1150.

Athena Chapekis and Emma Lieb. Google users are less likely to click on links when an AI summary appears in the results. Technical report, Pew Research Center, July 2025.

Tong Chen, Hongwei Wang, Sihao Chen, et al. Dense X retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*, 2023.

Lesley Chiou and Catherine Tucker. Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy*, 26(4):782–805, 2017. doi: 10.1111/jems.12207.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2409.00729.

Hannes Datta, George Knox, and Bart J. Bronnenberg. Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science*, 37 (1):5–21, 2018. doi: 10.1287/mksc.2017.1051.

Tim Frank. Building toward a sustainable content economy for the agentic web. Microsoft Advertising Blog, February 2026. URL https://about.ads.microsoft.com/en/blog/post/february-2026/building-toward-a-sustainable-content-economy-for-the-agentic-web. Accessed 2026-07-26.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *EMNLP*, 2023. arXiv:2305.14627.

Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010. doi: 10.3982/ECTA7195.

Matthew Gentzkow, Jesse M. Shapiro, and Daniel F. Stone. Media bias in the marketplace: Theory. In *Handbook of Media Economics*, volume 1, pages 623–645. Elsevier, 2015. doi: 10.1016/B978-0-444-63685-0.00014-0.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2019. arXiv:1904.02868.

Roni Goldshmidt and Miriam Horovicz. TokenSHAP: Interpreting large language models with Monte Carlo Shapley value estimation. *arXiv preprint arXiv:2407.10114*, 2024.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, 2019. doi: 10.1126/science.aau2706.

Andrew M. Guess, Brendan Nyhan, and Jason Reifler. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5):472–480, 2020. doi: 10.1038/s41562-020-0833-x.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 2023. SSRN 4404340.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *International Conference on Machine Learning*, 2022. arXiv:2202.00622.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos. Towards efficient data valuation based on the Shapley value. In *AISTATS*, 2019. arXiv:1902.10275.

Charles I. Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, 2020. doi: 10.1257/aer.20191330.

Esra Kayabali. Aws waf adds ai traffic monetization capability to help content owners charge ai bots for content access. AWS News Blog, June 2026. URL https://aws.amazon.com/blogs/aws/aws-waf-adds-ai-traffic-monetization-capability-to-help-content-owners-charge-ai-bots-for- Accessed 2026-08-23.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, et al. Deduplicating training data makes language models better. In *ACL*, 2022. doi: 10.18653/v1/2022.acl-long.577.

Mark A. Lemley and Bryan Casey. Fair learning. *Texas Law Review*, 99:743, 2021. SSRN 3528447.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023.

Shayne Longpre, Robert Mahari, Ariel Lee, et al. Consent in crisis: The rapid decline of the AI data commons. *arXiv preprint arXiv:2407.14933*, 2024.

Sendhil Mullainathan and Andrei Shleifer. The market for news. *American Economic Review*, 95 (4):1031–1053, 2005. doi: 10.1257/0002828054825619.

Ikhtiyor Nematov, Dimitris Trivedi, et al. Source attribution in retrieval-augmented generation. *arXiv preprint arXiv:2507.04480*, 2025.

Nic Newman, Richard Fletcher, Craig T. Robertson, et al. Reuters institute digital news report 2025. Technical report, Reuters Institute for the Study of Journalism, 2025.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, 2023. arXiv:2303.14186.

Andrea Prat and David Strömberg. The political economy of mass media. In *Advances in Economics and Econometrics*. Cambridge University Press, 2013. doi: 10.1017/CBO9781139060028.004.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *EMNLP*, 2024. doi: 10.18653/v1/2024.emnlp-main.347.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, et al. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*, 2021.

Matthew Sag. Copyright safety for generative AI. *Houston Law Review*, 2023. SSRN 4438593.

Jiachen T. Wang, Zhun Deng, Hiroaki Chiba-Okabe, Boaz Barak, and Weijie J. Su. An economic solution to copyright challenges of generative AI. *arXiv preprint arXiv:2404.13964*, 2024.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data Shapley in one training run. In *International Conference on Learning Representations*, 2025. arXiv:2406.11011.

A Proofs

Proposition 1

Fix a block b with covering set P_b and let π be uniform over orderings of the publishers. Under linear u, the increment in (2) contributed by block b is $1/n$ for whichever member of P_b appears earliest in π and zero for all others. By exchangeability each member of P_b is earliest with probability $1/|P_b|$, so the expected credit from b is $1/(n|P_b|)$ for each $p \in P_b$ and zero otherwise. Summing over blocks and using linearity of expectation gives $\varphi_p = \frac{1}{n}\sum_b \mathbb{1}\{p \in P_b\}/|P_b| = \varphi_p^{\text{EQUAL}}$. □

Proposition 2

Fix a block b and let π place the publishers in the order in which they published. Under linear u the increment in (2) contributed by b is $1/n$ for whichever member of P_b comes first in π and zero for the rest, exactly as in Proposition 1; the only change is that π is now a point mass rather than uniform, so no expectation is taken. The earliest member of P_b under a chronological π is f_b, the first publisher of b. Summing over blocks, $\varphi_p = \frac{1}{n}\sum_b \mathbb{1}\{f_b = p\} = \varphi_p^{\text{PROVENANCE}}$. □

Proposition 3

The coverage fraction $c_p(\pi)/n$ takes values in the finite set $\{0, 1/n, \ldots, 1\}$, on which $x^\alpha \to \mathbb{1}\{x = 1\}$ pointwise as $\alpha \to \infty$. The increments in (2) lie in $[0, 1]$ and π ranges over a finite set, so the limit passes through the expectation. In the limit the increment for p under π is $\mathbb{1}\{c_p(\pi) + a_p(\pi) = n\} - \mathbb{1}\{c_p(\pi) = n\}$, which is one exactly when the publishers ahead of p in π leave some block uncovered and those up to and including p cover every block. That is the marginal contribution of p to v_{all} at the predecessor set induced by π, and averaging it over uniform π is the definition of the Shapley value of v_{all}, which is $\varphi^{\text{COMPLETER}}$. □

Proposition 4

Let $v_b(S) = \mathbb{1}\{S \cap P_b \neq \emptyset\}$, so that $v_{\text{lin}} = \frac{1}{n}\sum_b v_b$. Fix b. Any $p \notin P_b$ satisfies $v_b(S \cup \{p\}) = v_b(S)$ for all S and is therefore a null player, receiving zero. Any two $p, q \in P_b$ are interchangeable in v_b, so by symmetry they receive equal amounts. Efficiency requires the payments to sum to $v_b(\mathcal{P}) = 1$. Hence each $p \in P_b$ receives $1/|P_b|$. The Shapley value is additive in the game, so the value of v_{lin} is $\frac{1}{n}\sum_b$ of these, which is exactly φ^{EQUAL}. □

Proposition 5

Write $v_{\text{prov}} = \frac{1}{n}\sum_b w_b$ with $w_b(S) = \mathbb{1}\{f_b \in S\}$. Each w_b is a dictator game: f_b determines the value on its own and every other publisher is a null player, so efficiency and the null-player

axiom force the Shapley value of w_b to pay 1 to f_b and 0 to everyone else. Additivity then gives $\varphi_p = \frac{1}{n}\sum_b \mathbb{1}\{f_b = p\} = \varphi_p^{\text{PROVENANCE}}$. □

Remark 2. The symmetry axiom is what separates the two representations. In v_{lin} the members of P_b are interchangeable, so any symmetric value must split b equally among them and PROVENANCE is unreachable; PROVENANCE is recovered either by dropping symmetry, which is what a non-uniform π does, or by moving to a game in which the publishers of b are no longer interchangeable, which is what v_{prov} does. Timestamps enter the mechanism at exactly one point, and this is it.

Remark 3. Note that $v_{\text{all}} = \min_b v_b$. Unlike v_{lin} it is not additively separable across blocks, which is why COMPLETER admits no closed form and must be evaluated by sampling permutations.

B Robustness

B.1 The premium classification

The premium tier is the least mechanical construct in the paper, so we report the headline under four alternative definitions. The concern is that a hand-assigned prestige coding might be doing the work that we attribute to priority.

Table 7: Premium share under alternative definitions of the tier, global index.

Definition	Publishers	EQUAL	PROVENANCE	Spread
Pooled premium (headline)	137	41.87%	49.55%	7.68
Type-based newsroom coding	149	43.61%	51.36%	7.75
`premium-us` only	54	20.09%	28.70%	8.61
`premium-intl` only	83	21.79%	20.85%	-0.94

The type-based coding marks a publisher as a newsroom if its independently assigned *type* is one of national news, broadcast news, wire service, business/financial, magazine/long-form or digital news. It makes no reference to prestige and was not constructed for this exercise. Spread is PROVENANCE − EQUAL in percentage points.

The first two rows are the reassuring part. Replacing the prestige tier with the independent type coding moves the levels by under two points and the spread by less than a tenth of a point, from 7.68 to 7.75. The result is a property of which publishers report first, not of how we labeled them.

The last two rows are not reassuring, and we think they are the more interesting finding. The entire provenance premium accrues to American premium outlets, whose share rises from 20.09% to 28.70%, a spread of 8.61 points. International premium outlets gain nothing: their share *falls* slightly, from 21.79% to 20.85%, a spread of -0.94 points. **So Provenance does not reward premium publishers as a class. It rewards American premium publishers, and it is close to neutral between international premium outlets and the standard tier.**

We do not want to over-read this from a single episode, and there is an obvious mechanical candidate for at least part of it: the corpus was crawled from a US-facing product, and the

Table 8: Extraction- and clustering-error robustness, global index.

Perturbation	Rate	EQUAL	PROVENANCE	Spread
none	—	41.87%	49.55%	7.68
Misfiled kind	0.05	41.82%	49.56%	+7.74
	0.20	41.66%	49.28%	+7.63
Wrongly split	0.05	41.81%	49.46%	+7.65
	0.20	41.55%	49.34%	+7.79
Wrongly merged	0.05	41.89%	49.66%	+7.77
	0.20	41.97%	50.08%	+8.12

Mean of five replicates. Spread is PROVENANCE − EQUAL in percentage points. A rate of 0.20 means one publisher record in five is wrong, which is far beyond any error rate we would defend for this pipeline; it is included to show where the result would have to break.

timestamp quality reported in Section 5.1.2 may differ systematically by region, so non-US outlets may be losing ties they in fact won. But it is exactly the kind of incidence a marketplace designer needs to see before adopting a first-mover rule, and it is invisible in the pooled number.

B.2 Extraction and clustering error

The index is produced by a language model, so it is wrong some of the time. We do not have ground truth against which to measure how often, so instead we ask how much error the headline could absorb. Three error modes move the covering set P_b of a cited block in different directions, and we simulate each directly on the index.

Misfiled kind A block is recorded under the wrong kind. The publisher leaves b and lands on the block with the same cluster identifier and the other kind, while publishers of that sibling leak into b. The size of P_b is roughly preserved and its composition is churned.

Wrongly split The clusterer fails to see that two publishers said the same thing, so a publisher is detached into a cluster of its own. That cluster is not the one the answer cited, so the publisher is not paid; $|P_b|$ shrinks.

Wrongly merged The clusterer conflates two distinct claims, so publishers who do not carry b are recorded as carrying it; $|P_b|$ grows.

Each mode is applied independently to every publisher of every cited block at rate ε, and EQUAL and PROVENANCE are recomputed on the perturbed index. Both rules are closed form per block, so no sampling is involved; we average over five replicates. At $\varepsilon = 0$ the procedure reproduces the numbers of Section 5.2.2 exactly, which is the check that it is measuring what we intend.

Table 8 reports the result, and it is dull in the way one wants a robustness table to be dull. Across all three modes and rates up to 20%, the spread stays between +7.62 and +8.12 points against a baseline of 7.68. The reason is structural rather than lucky. Splitting and misfiling both thin the covering sets—mean $|P_b|$ falls from 22.4 to about 19 at the highest rate—but they thin the sets of premium and non-premium publishers in the same proportion, and both rules are homogeneous of degree zero in a uniform thinning. Only merging moves the spread at all, and it moves it *up*: conflating claims imports extra publishers into a block, which dilutes the equal split faster than it dilutes a first-mover rule whose winner is usually already the earliest publisher in the enlarged set.

The one caveat worth stating is that our perturbations are uniform, and real extraction error will not be. If the model is systematically worse at parsing one class of outlet—shorter wire copy, non-English sources, paywalled excerpts—the error is correlated with the very tier distinction the headline is about, and no amount of uniform perturbation would reveal it. Measuring that requires labeled ground truth, which we do not have.

B.3 Order and utility

Section 5.2.4 reports the article-level against publisher-level randomization and the Monte Carlo validation. To summarize: the two randomizations differ by 5.57 percentage points of total variation in the global index and 0.07 in the local index, and the sampled $\alpha = 1$ arm reproduces the closed-form EQUAL payout vector to 0.57 points, which is the sampling error implied by 200 permutations per query and bounds the precision of every sampled quantity we report.

B.4 Monte Carlo convergence

Every α-indexed quantity in this paper is estimated from 200 publisher orderings per query. This section asks what that number buys, and it can ask the question unusually sharply because at $\alpha = 1$ the estimand is known in closed form: by Proposition 4 the sampler is estimating the equal per-block split, which we can also compute exactly. So we can measure error against ground truth rather than only against the sampler itself.

We run three independent streams of 800 orderings per query and read the estimate off prefixes of each stream, which yields every sample size from one pass. Two quantities are reported, and they answer different questions. *Accuracy* is the total variation between the sampled credit vector and the exact one, available only at $\alpha = 1$. *Precision* is the mean pairwise total variation between the independent streams, which is defined for every rule.

Figure 12 shows both. The accuracy curve falls at a fitted rate of $m^{-0.503}$, and the precision curves at rates between $m^{-0.497}$ and $m^{-0.517}$: the textbook $m^{-1/2}$, with no sign of a slower regime. Concretely, the distance from the truth falls from 1.59 points at $m = 25$ to 0.56 at $m = 200$ and 0.28 at $m = 800$.

The two measures are related in a way that lets us export the closed-form benchmark to the rules that have none. If two independent estimates carry independent errors of comparable size, the distance between them should exceed the distance from either to the truth by a factor of $\sqrt{2}$. At $\alpha = 1$ the observed ratio is 1.40, against $\sqrt{2} \approx 1.414$. Taking that seriously, the between-stream distances at $m = 200$ imply errors of roughly 0.50 points at $\alpha = 2$, 0.74 at $\alpha = 8$ and 0.99 for the completer. The completer is the hardest of the family to estimate, which is what one expects of a rule that depends on the identity of a single pivotal publisher rather than on a smooth average.

These magnitudes are what license the readings in Section 5.2.4. The gap of 0.57 points between the closed-form EQUAL rule and its sampled counterpart is squarely inside the sampling error and should be read as noise, which is the point of reporting it. By contrast the distances that carry the paper's argument—19.6 points from EQUAL to the completer, 33.7 points from PROVENANCE to the completer—are more than an order of magnitude larger than the noise floor, so no plausible increase in m would disturb them.

One asymmetry is worth flagging because it recurs elsewhere in the paper. The *aggregate* converges far faster than the *allocation*. At $m = 200$ the premium share is within 0.011 points of its exact value while the credit vector is still 0.56 points away in total variation, a factor of about fifty; the premium share is already accurate to a twentieth of a point at $m = 25$. Averaging over a thousand publishers cancels most of the sampling error, but the individual payments it cancels are exactly what a marketplace would actually pay out. A settlement system should therefore not choose m by watching the headline. The same caution appears substantively in Section 5.2.4, where PROVENANCE and the completer agree on the aggregate to within a hundredth of a point while differing by 33.7 points on who gets paid.

C Discussion: Global indexing

In Section 3 we described our system for answering queries and attributing credit using the global index. We now return to our choice to use a global, pre-computed index and compare it with alternative solutions.

C.1 Index scope: local and global

Recall from Section 3.4 that, for any given query, a search engine is used by a retriever system to return a small grounding set of articles. Every rule in Section 4 is defined with respect to a choice of covering sets P_b for each block b extracted from the grounding set. The choice of P_b requires deciding which corpus to consult, and there are two coherent answers.

The global index. P_b is every publisher in the index who carried block b, whether or not their article was retrieved for this query. The counterfactual is one of *publication*: what would the answer have looked like if this fact had never been reported at all?

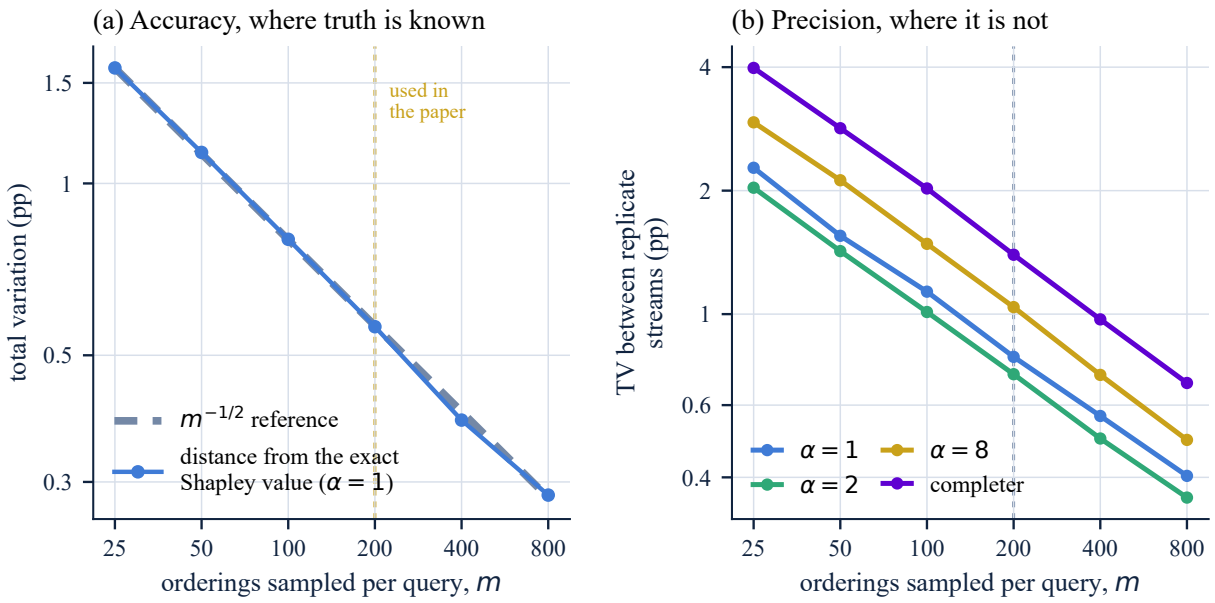


Figure 12: Monte Carlo convergence. Panel (a): distance from the exact Shapley value at $\alpha = 1$, against a $m^{-1/2}$ reference the data are indistinguishable from. Panel (b): mean pairwise distance between three independent streams, for rules with no closed form. The dashed vertical line marks the $m = 200$ used throughout the paper.

The local index. P_b is every publisher who carried block b *in the query's grounding set*—the handful of articles the retriever actually surfaced. The counterfactual is one of *retrieval*: what would the answer have looked like if we had not shown this particular article during grounding?[6]

C.2 Attribution timing: Index time and inference time

A second, related aspect of our proposed appproach concerns *when* attribution is computed.

Inference-time attribution waits for a query and, while generating an answer, works out which of the retrieved documents caused which parts of the generated responses. Three existing approaches to credit attribution are performed at inference time: asking the model to cite its sources (Gao et al., 2023; Bohnet et al., 2022); perturbing the context and re-running, as in ContextCite (Cohen-Wang et al., 2024) or Shapley over retrieved documents (Nematov et al., 2025); and reading the model's internals (Qi et al., 2024).

Index-time attribution, which our system employs, calculates attribution as part of the indexing work that occurs before any query exists. Each article is decomposed into blocks and merged into the index once, when it is crawled, so every block already carries the list of outlets that published it and the time each of them filed. When a query arrives, settlement simply reads

[6]The terms collide with an established usage in distributed information retrieval, where "local" and "global" index refer to document- and term-partitioned shards of a search index. We use them in the economic sense throughout: the scope of the counterfactual, not the physical layout of the index.

the covering sets off the index and divides the fee. This is the pipeline described in Section 3.

Timing and scope are different distinctions, but they are not independent. Working out attribution at inference time effectively forces the use of the local index, because it is impractical to recalculate the global index for every query. When building a global index in advance, one could still use the local index for attribution. We argue below, however, that using the global index has substantial advantages for accurate credit attribution.

C.3 The computational cost of index-time and inference-time attribution

A practical disadvantage of index-time computation the computational cost of constructing the index in advance. Indeed, indexing is often considered to be expensive compared to inference-time calculation. We note, however, that index construction is an up-front cost that makes both answer generation and credit attribution—both per-query costs—less expensive. Moreover, in our system, the covering set of articles is already materialized in the index, so settlement of credit and payments reduces to set arithmetic. That is also what makes the *global* index relatively affordable. We therefore expect a global index with index-time attribution to be superior when the number of queries is sufficiently large, as should be the case for a content marketplace that is large enough to meaningfully influence publisher outcomes.

C.4 Further arguments in favor of index-time computation

We want to be fair to the inference-time literature: those methods are excellent at what they are designed for, which is explaining and verifying a generation. They are the wrong instrument for paying publishers, for five reasons beyond cost.

The first and most important is that they answer the wrong counterfactual. Inference-time methods ask which pieces of *this* context moved the output. The marketplace question is what a publisher is worth given everything else that exists. These differ whenever a fact appears in more than one source, which in our data is the common case. A perturbation method that removes one outlet's document watches the answer degrade and pays that outlet; it cannot see that three other outlets reported the same event and any of them would have served. Section 5.2 shows that inside the local index every rule in the family returns the same answer to within one percentage point, so this is not a matter of precision: **inference-time attribution cannot tell a fact only one outlet had from a fact twenty outlets had, even in principle, because the other nineteen were discarded before the query arrived.**

Second, they launder retrieval bias into payment. If the retriever is popularity- or SEO-driven, whoever is retrieved is paid, and being paid funds being retrieved. Inference-time attribution has no defense, because retrieval is upstream of it. A global covering set at least allows a rule to be defined over all publishers who reported the fact rather than the k the retriever surfaced. We note honestly in Section 5.3 that our own estimates are still computed over *cited* blocks and so still condition on retrieval; the framework does not have to.

Third, they are not auditable, and markets require auditability. A publisher cannot verify a saliency map or a perturbation run: it cannot see the prompt, cannot re-run the model, and cannot reproduce a sampled generation. A block index is a persistent, inspectable artifact—these are your blocks, this is who else carried them, this is when you filed—against which disputes can be resolved. Contracts must be written on verifiable states, and an unauditable metering standard is not a metering standard.

Fourth, they are unstable and gameable. Generation is stochastic and self-citation is unreliable: Liu et al. (2023) find only 51.5% of sentences fully supported by their cited sources. Paying on a signal with that error rate is untenable. Duplication attacks—republishing content so that more chunks are retrieved—inflate credit under any per-retrieved-chunk scheme, whereas cross-publisher deduplication neutralizes them by construction, since a block is credited once however many documents carry it.

Fifth, provenance is impossible at inference time. First-mover credit requires knowing who reported a fact first. That is a property of the corpus's history, not of a single prompt, and no amount of context attribution recovers it. Any marketplace that wishes to reward original reporting must therefore maintain something like a block index.

C.5 Counter-arguments to the global index

Three concessions are due. First, a global index is susceptible to *extraction error*: index-time attribution inherits whatever mistakes the decomposition and clustering make, and inference-time methods have no such dependency. We investigate these errors in our empirical exercise via robustness checks in Section 5.3. Second, there is a non-trivial cost of *maintaining the index*. This is a recurring bill: fresh content must be indexed before it can be attributed, which is a latency the inference-time approach does not have. As we argued above, this cost is per-document and may be quite low when amortized over all queries. But in a setting that requires serving few queries over a large and fast-churning corpus, our global indexing method may not be appropriate. Third, *the two methods are complementary*: even when index-time attribution is effective for credit attribution, inference-time attribution remains the right tool for verifying that a model actually used what it cites. A mature marketplace would plausibly use inference-time methods to police the citation step and index-time methods to settle attribution.