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

Daniel G. Goldstein[1], Jake M. Hofman[1], Nicole Immorlica[1], Oliver Kiss[2], Brendan Lucier[1], Ishan Mahajan[3], Markus Mobius[1], Jingyi Qiu[4], Tanya Rosenblat[4], David 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*

August 23, 2026

Working paper — preliminary and incomplete — Do not distribute

Abstract

We propose and evaluate a block-level publisher content marketplace: a settlement layer that attributes AI-generated answers not to entire web pages, but to deduplicated "lego blocks" of news and commentary held in a shared index. We show that three intuitive payment rules—equal split for shared facts, first-mover reward, and indispensability—all emerge naturally from a simple news-reading model, and that each can be mapped to a Shapley value. This framework is governed by two behavioral dials (reading order and choice of utility function) and one structural dial: the index scope. Evaluating across a set of 1,974 real-world grounded user queries reveals a stark contrast. Against a *local index* (the limited search results used by the LLM to generate an answer) most cited blocks appear unique, forcing every payment rule to yield identical payouts. Only against a *global index*, the complete universe of all published news, does source overlap become visible, allowing payment rules to diverge and significantly shift premium publishers' share of value from 41.87% under an equal split to 46.3–49.6% under a first-mover rule. Because inference-time attribution sees only the local search window, it is fundamentally incapable of expressing any attribution policy: the information that separates the rules was discarded before it ran. Attribution must be computed against a global index, before any query arrives.

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

*Authors are listed alphabetically. The lego decomposition approach was developed with Luca Braghieri, Sarah Eichmeyer and Ro'ee Levy, together with Kiss, 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

For thirty years the web paid publishers in clicks. A search engine or a social feed showed a headline, the reader traveled to the publisher's page, and the publisher monetized the visit. The bargain was never comfortable, but it was clear enough: the intermediary sent traffic to publishers' sites while diverting some of the reader's attention—and the advertising revenue that follows it—to itself.

Generative search breaks this bargain. When an assistant answers a question directly, it consumes the publisher's reporting while returning almost no traffic. Chapekis and Lieb (2025) find that users click through on 8% of searches carrying an AI summary (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 an urgent market-design problem that platforms are actively attempting to solve. 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 content actually used, it leaves the underlying *mechanism* open. Delivered value must be measured, which poses a fundamental accounting challenge.

The difficulty begins with the unit of account. Media consumption has unbundled over two decades: from the newspaper, to the article, and now to the individual *fact*. Music underwent the same shift from album to track to playlist, re-pricing the underlying media at every step. An AI answer is synthesized from facts drawn from many articles simultaneously; it is not a pointer to a single page. If a marketplace's unit of account remains the page, it measures something the product no longer consumes.

Facts, however, are awkward to price. They are *non-rival* (one reader's consumption does not deplete the fact) (Jones and Tonetti, 2020) and *non-excludable* (once published, anyone can 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. Crucially, an index alone cannot tell the two apart. Any payment rule must therefore take an explicit stance on whether filing first deserves financial reward.

This paper makes four core contributions:

Contribution 1: A unit of account for news. We propose a fine-grained, interpretable unit of account. Each crawled article is decomposed by a language model into atomic *"lego" blocks*: *news* blocks record state-changing actions in the world, while *commentary* blocks record stances, opinions, or context regarding those actions. Identical blocks across publishers are clustered into a shared, tagged index that records every atomic claim, its publishers, and their respective timestamps.

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*: twenty distinct articles reporting the same event resolve to a single block with twenty publishers, collapsing redundant coverage onto a single numeraire.

As Sections 3 and 4 demonstrate, building these covering sets at index time allows queries to be settled via straightforward table lookups. Furthermore, generating answers from deduplicated blocks significantly reduces token overhead by inserting shared facts into the model's context window once rather than repeatedly.

Contribution 2: A unifying economic theory. Once articles map to lego blocks, the payment problem simplifies: for every fact used in an answer, the index identifies who reported it and when. We evaluate three intuitive payment rules drawn from existing literature:

1. Equal division among all publishers that carried the fact (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).

Section 5 shows that all three rules emerge naturally from a simple model of news reading, and that each can be mapped to a Shapley value. 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 by the facts it introduces first.

We prove that setting the reading order to *random* and utility to *linear* yields EQUAL; setting the order to *chronological* yields PROVENANCE; and holding the order random while pushing utility toward an *all-or-nothing* threshold yields COMPLETER. This framework provides a structured foundation for evaluating payment mechanisms, mapping them as points along two

[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).

parameterized economic dimensions: the *order dial* (whether speed deserves reward) and the *utility dial* (how heavily completeness is valued).[2]

Contribution 3: Identifying the dominant dial. Section 6 introduces the parameter that dominates all others: the index scope over which attribution is measured. Under a *global index* (the complete corpus of published news), a publisher's contribution is evaluated against every outlet that reported the fact. Under a *local index* (the limited search results used by the LLM to generate an answer to a specific query), compensation is evaluated only against the temporary retrieval window.

Contribution 4: Empirical findings and index-time necessity. The empirical divergence between these two scopes is stark. Against a global index, only 27.2% of cited blocks have a single publisher. 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 (with a mean coverage of 1.23), causing every payment rule in the family to collapse to the same payout within one percentage point. Attribution rules matter only when duplication is visible, and duplication is only visible globally.

This insight exposes a critical flaw in current system designs. A rapidly growing literature computes attribution at *inference time* by perturbing or inspecting retrieved context post-query (Cohen-Wang et al., 2024; Qi et al., 2024; Nematov et al., 2025). While effective for answer *verification*, inference-time methods are fundamentally incapable of executing content *compensation*: they operate strictly within the local retrieval index, where all policy rules yield identical payouts. Inference-time approaches cannot distinguish an originator from a copier because that global context was discarded prior to query execution.

Section 6 provides the cost accounting behind this trade-off. While inference-time methods avoid initial setup costs, they incur ongoing computation costs on every query without the capacity to express true attribution policies. Index-time computation pays a higher one-time indexing cost per document, but reduces query-time generation costs and renders policy-based attribution computable at zero cost per query.

The paper proceeds as follows. Section 2 reviews related literature. Sections 3 and 4 detail index construction and answer generation. Section 5 develops the unified Shapley mechanism. Section 6 demonstrates why attribution must be evaluated against a global index at index time. Section 7 describes our datasets, Section 8 presents our empirical results, Section 9 analyzes publisher incentives, and Section 10 outlines current limitations.

[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. Its emergence is a by-product of the two dials rather than a proposal.

2 Related literature

The paper sits at the intersection of six distinct literatures, synthesizing them into a unified market-design framework. We take them in the order in which they bear on our core argument: starting with the economic structure of news reporting, progressing through referral history and product unbundling, grounding in the economics of information goods, and concluding with technical data valuation, generative attribution, and legal frameworks.

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 the extremes of our order dial. As detailed in Section 9, streaming also offers a cautionary tale: rigid allocation rules often invite strategic manipulation that market designers fail to anticipate.

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). This contextual shift is fundamental: in training, the coalition consists of a vast corpus, making exact Shapley values intractable and forcing reliance on approximations. In serving, the coalition comprises only the handful of publishers cited in a given answer, rendering the Shapley value exact and computable (Section 5).

Attribution vs. Settlement in Generative AI. Generative attribution literature serves as the primary contrast class for Section 6. 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 6 and quantify in Section 8. 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.

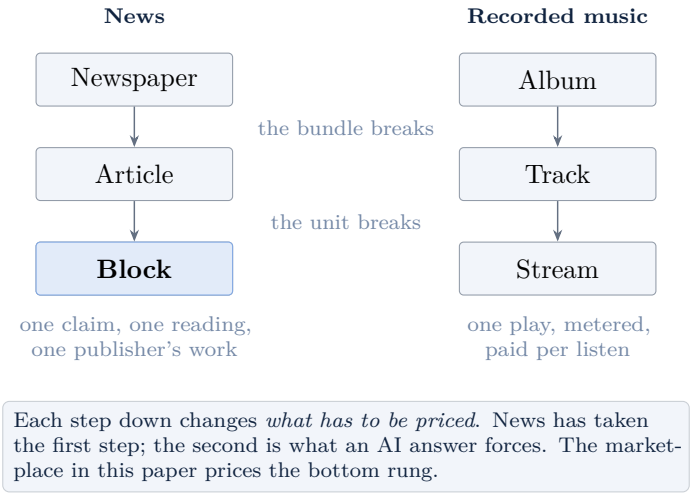


Figure 1: The unbundling ladder. Newspaper to article to block, beside the album–track–stream analogue. Each rung changes the object that must be priced.

3 From articles to blocks: the lego index

This section describes the settlement layer of our proposed marketplace. The index is constructed once across the entire web crawl and contains no information beyond what is present in those source articles. Rather, it offers a distinct representation of the corpus designed specifically to make economic attribution computationally efficient. Where a crawl is a collection of unstructured documents, 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 deterministic table lookup.

3.1 The unbundling ladder

Media consumption has undergone two structural unbundlings. The newspaper served as the unit of subscription; search and social platforms shifted the unit of attention to the individual article; generative AI now turns the individual claim into the unit of use. Music followed an identical trajectory—from album to track to playlist—with each transition altering the fundamental unit to be priced (Figure 1). Because an AI answer synthesizes claims drawn from many pages simultaneously rather than directing readers to a single URL, a marketplace tethered to page-level accounting meters an asset the underlying product no longer consumes.

3.2 Decomposition and the block tree

Each crawled article undergoes a sequence of language-model extraction passes that decompose the text 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 indi-

viduals, 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 carried across multiple 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.

Crucially, extracted blocks form 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). This tree structure is essential for deduplication. 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. Deduplication clustering therefore evaluates a commentary block together with its parent news block. Section 4 demonstrates how this tree representation serves a second vital role during answer generation.

Separating news from commentary is not decorative; it isolates the two economically distinct goods a publisher sells— being first to a fact versus providing a perspective or context on it—which perform markedly differently under different payment rules (Section 8). Figure 2 illustrates a worked example.

3.3 Clustering across publishers

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 step constructs the covering sets P_b analyzed in Section 5.

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. The index thus answers three fundamental questions about any atomic claim: who reported it, who filed it first, and how many alternatives exist across the market. Figure 3 traces this ingestion process: shared claims stack additional publisher tags on existing rows, while novel claims append new rows.

We will also refer to the *local index* as the set of deduplicated blocks that are constructed from the subset of articles retrieved for constructing the answer to a given query.

3.4 Properties, scope, and limitations

The global index is inspectable: publishers can directly inspect their indexed lego blocks, verify timestamp records, and audit the set of co-reporting publishers. Section 6 argues that this

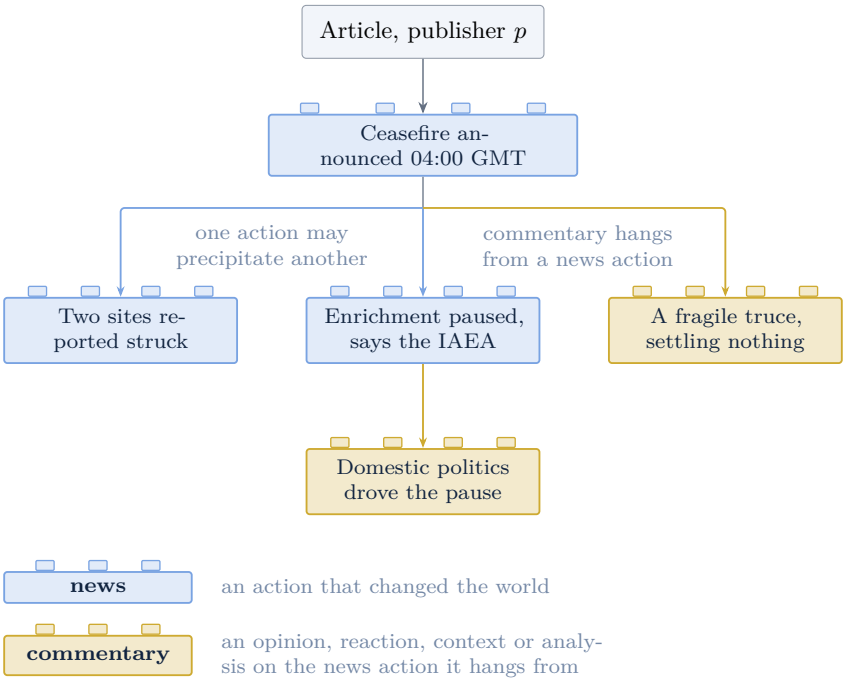


Figure 2: A block tree. One article decomposed into news and commentary blocks. Commentary blocks hang from the news block they are about, and a news block may hang from an earlier action that precipitated it. The decomposition takes several language-model passes per document, and runs as soon as the search index crawler discovers a new article.

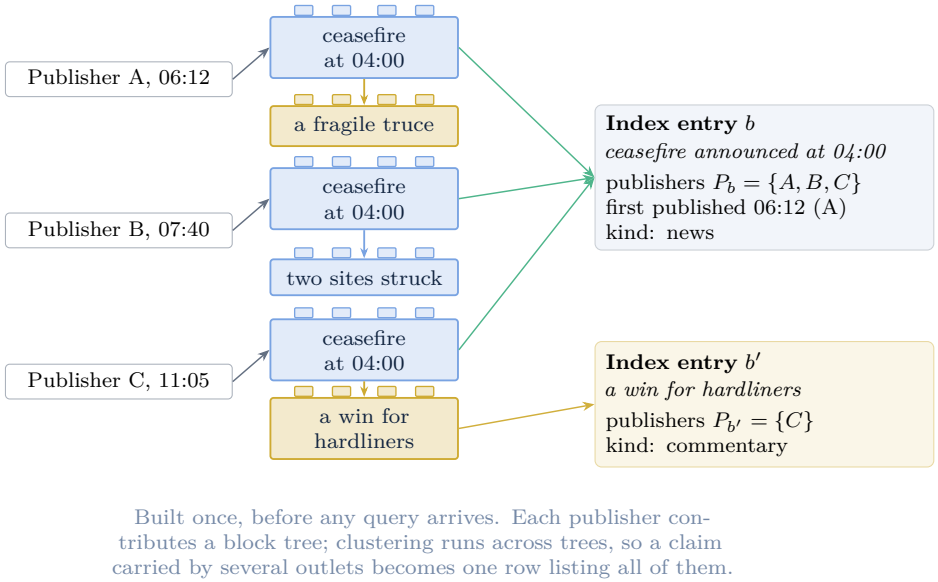


Figure 3: Building the index. Each publisher contributes a block tree; blocks from separate trees that express the same claim collapse into a single index row. The row records who carries the claim and who filed first.

transparency is a prerequisite for a functional content marketplace.

The global index is simply a different representation of the index of all news URLs that is optimized for attribution. Without this index, it would have to be constructed at query time, which imposes a huge computational cost. For this reason inference-time attribution methods restrict their view to the subset of URLs retrieved by a search engine that is used for generating the answer (what is called the "grounding" of a response). Section 6 details the economic costs of that constraint.

The index is bounded by the reach of the web crawl (the directional bias of which is evaluated in Section 10). Furthermore, because decomposition relies on language models, it inherits potential extraction errors. Lacking labeled ground-truth data to compute absolute error rates, Appendix B.2 conducts sensitivity analyses by perturbing the index across all error modes at rates up to 20%. The core empirical conclusions remain robust under these perturbations.

Finally, the index measures *coverage*, not *veracity*. It records that an outlet published a claim, not whether the 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.

Figure 4 illustrates the resulting distribution of cluster sizes across news and commentary blocks for all extracted claims versus those cited in generated answers.

4 Answering a query, and splitting the credit

4.1 From query to answer

When a query arrives, a search engine retrieves a grounding set of relevant articles. 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 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 5 illustrates this query pipeline.

Because the model explicitly outputs cited block identifiers, attribution requires no additional post-hoc LLM calls or context perturbations. Settlement reduces to a deterministic lookup against the indexed covering sets.

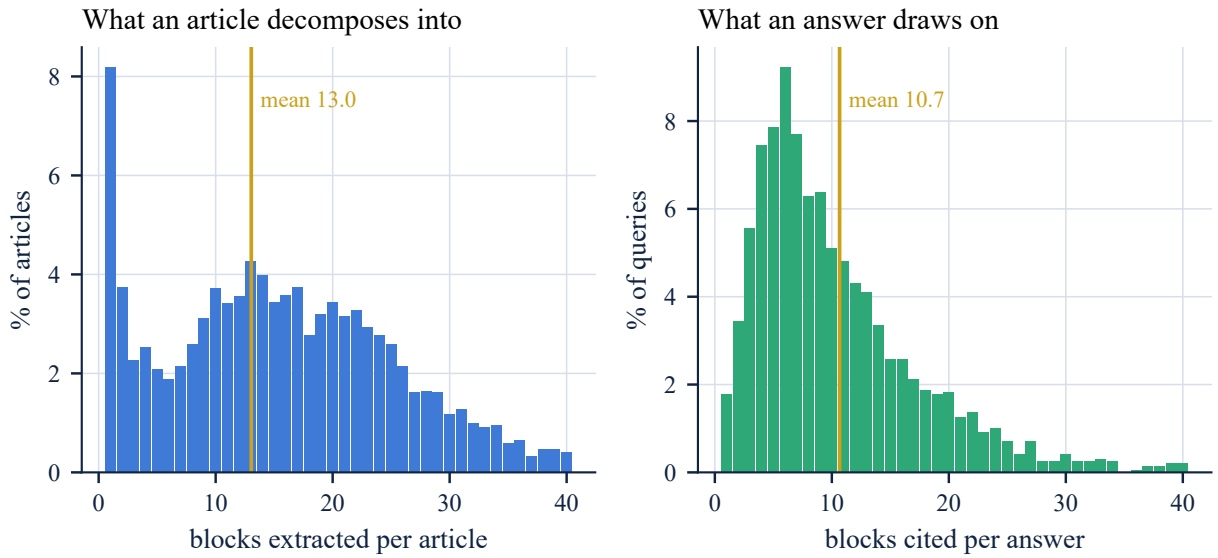


Figure 4: Anatomy of the index. Left: blocks extracted per article. Right: blocks cited per answer. Both truncated at 40 for legibility.

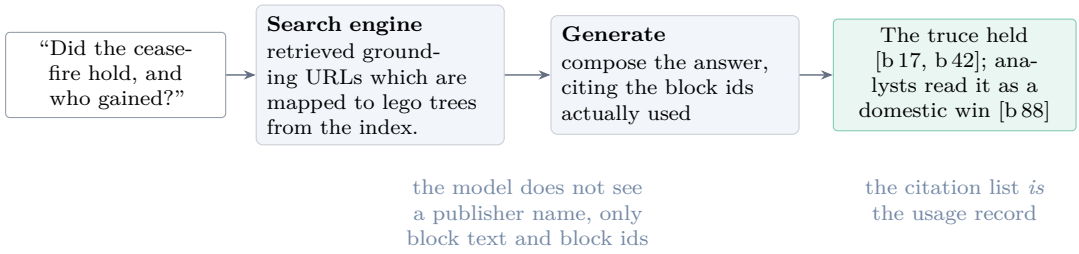


Figure 5: Answering. A query arrives, deduplicated block trees rather than raw article text are placed into context, and the model cites specific block identifiers in its response.

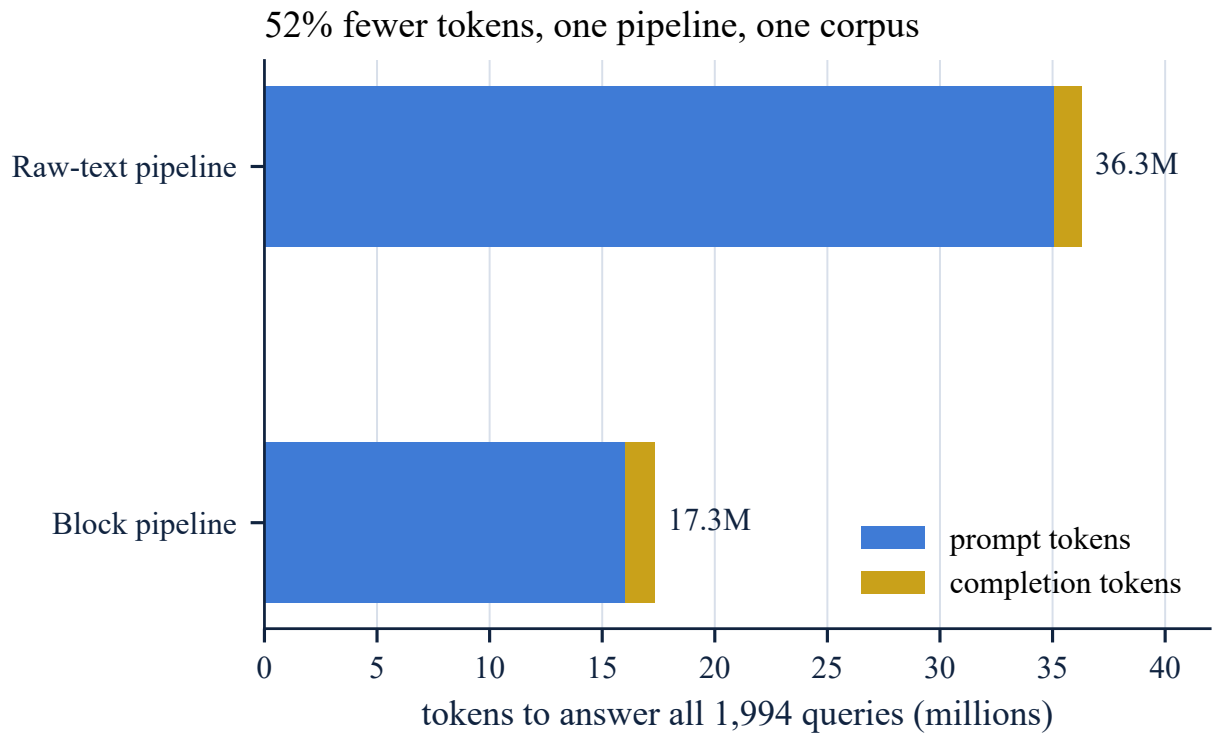


Figure 6: The answering cost (C_2). Prompt and completion tokens per query, comparing deduplicated block trees against raw article text across 1,994 queries.

4.2 Impact on generation costs

In addition to enabling transparent attribution, deduplicating facts before insertion into the prompt yields a significant computational side effect: it substantially reduces 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).

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 return to this full accounting ledger in Section 6.3.

4.3 The attribution problem

With the answer generated, the marketplace faces its core economic problem: a set of cited blocks must be attributed, and a unit of value assigned to each block must be distributed among the relevant publishers.

Consider the minimal non-trivial example illustrated in Figure 7. An answer cites three blocks: a news block reporting a breaking event, published first by outlet A at 09:14 and subsequently carried by outlets B and C; a second news block reported exclusively by B; and a commentary block offering a unique perspective, carried exclusively by A. How should credit be allocated

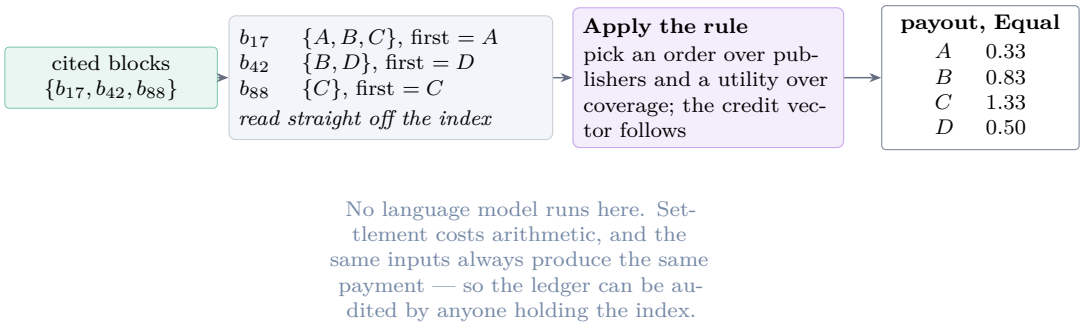


Figure 7: Attribution as table lookup. Cited block identifiers are matched against the pre-indexed database, which records covering sets and timestamps. Applying the payment rule is purely algorithmic and deterministic— reproducible by any party with access to the index without requiring further language model calls.

across these outlets? If value tracks broad coverage, A and B each receive full credit for their exclusive blocks, while credit for the shared news block is split equally three ways among A, B, and C. Publisher A receives $\frac{1}{3} + (\frac{1}{3} \cdot \frac{1}{3}) = \frac{4}{9}$, while C receives $\frac{1}{9}$ for providing a fact carried by two others.

If value tracks original reporting, C receives zero, and A captures the entirety of the shared news block for having scooped the story.

If value tracks indispensability, both A and B are rewarded for supplying unique content that could not be sourced elsewhere, while C again receives nothing—though for a distinct economic reason, yielding a different split between A and B.

These three approaches correspond directly to the payment rules evaluated in this paper. While they appear ad hoc when presented as distinct formulas, Section 5 shows that all three emerge from a single, unified model of news consumption, and that each can be mapped to a Shapley value.

5 One mechanism, three dials

Three payment rules suggest themselves, and they are normally presented as competing philosophies of desert: pay everyone who carried the fact, pay whoever broke it, or pay whoever supplied what nobody else could. This section shows that all three emerge from a single mechanism—a stylized model of news reading—and that the apparent philosophical disagreement is a disagreement about the settings of two dials, plus a third, developed in Section 6, which turns out to matter more than either. Section 5.5 then relates the settings to the Shapley value, which each of the three rules turns out to be, though not all of the same game. The names we give the rules are ours; the positions are not.

We present this as an instrument rather than as a theoretical contribution. Its purpose is to make the measurement in Section 8 principled: once the rules are points in a space, the numbers

we report are comparative statics along axes that a market designer can hold an opinion about, rather than the outputs of formulas chosen by precedent.

5.1 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.

Equal. Each block is worth $1/n$ and is split evenly among its publishers: $\varphi_p^{\text{EQUAL}} = \frac{1}{n} \sum_{b \in \mathcal{B}} \frac{\mathbb{1}\{p \in P_b\}}{|P_b|}$.

Provenance. Each block's value goes entirely to whoever published it first, with ties split evenly: $\varphi_p^{\text{PROVENANCE}} = \frac{1}{n} \sum_{b \in \mathcal{B}} \frac{\mathbb{1}\{p \in \arg\min_{q \in P_b} t_{q,b}\}}{|\arg\min_{q \in P_b} t_{q,b}|}$.

Completer. Each publisher is paid for being indispensable—for supplying what no combination of the others could. This is the rule computed in the applied literature (Ghorbani and Zou, 2019; Jia et al., 2019), and unlike the other two it is defined as a Shapley value rather than by a formula. Consider the cooperative game in which a coalition $S \subseteq \mathcal{P}$ is worth one if it can supply *every* block of the answer and nothing otherwise, $v_{\text{all}}(S) = \mathbb{1}\{S \cap P_b \neq \emptyset \text{ for all } b \in \mathcal{B}\}$, and let $\varphi^{\text{COMPLETER}}$ be its Shapley value. An answer that is missing anything is worth nothing, so a publisher earns only in those orders of arrival in which it is the one that completes the set. The rule has no closed form and we evaluate it by sampling permutations.

The first two are formulas and the third is a solution concept, which is part of why they are non-obvious to compare. The rest of this section puts all three on the same footing.

5.2 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 gets nothing at all unless the answer is completely

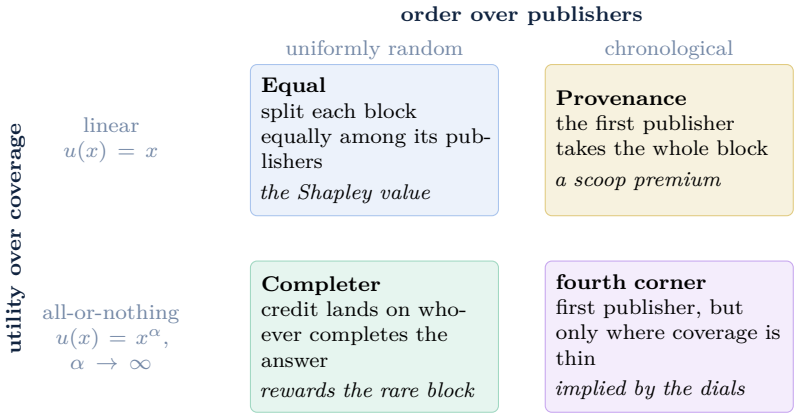


Figure 8: Two dials, four corners. The same cited blocks divided under each combination of reading order and utility curvature. Three of the corners are rules already in circulation; the fourth is the combination the two dials imply.

covered. The exponent therefore has a plain reading—*how badly does the buyer need the answer to be complete?*—and it is the second of the two dials.

Each publisher is paid the increment in the reader's satisfaction that it caused. Writing $c_p(\pi)$ for the number of blocks already covered when p is reached under order π, and $a_p(\pi)$ for the number of blocks p is the first to supply,

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

Equation (2) is the whole mechanism. It has exactly two free parameters: the distribution over reading orders π, and the exponent α. We take them in turn. Figure 8 illustrates the construction.

5.3 The order dial

The natural first guess is that reading in a random order rewards whoever got there first, and therefore reproduces Provenance. It does not. Under a uniformly random order, 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.

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

Proof sketch. 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, which is $\varphi^{\textsc{Provenance}}$. □

Proposition 2 is nearly trivial, and we state it only to make the comparison with Proposition 1 exact: the two rules differ in the distribution over π and in nothing else.

The economics of the dial is now visible, and it is not a technicality. A random order is a veil of ignorance: it says the market should not care who happened to be first, only who was able to supply the fact. Chronological order says the opposite—that being first is the thing being bought. This is a normative choice about whether original reporting or availability is what the marketplace exists to fund, and Section 8 shows it is worth several percentage points of aggregate revenue share.

Remark 1. 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.

5.4 The utility dial

Hold the order dial at random and turn the second one. 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.

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 $\alpha \to \infty$.*

Proof sketch. The coverage fraction takes values in the finite set $\{0, 1/n, \dots, 1\}$, on which x^α converges pointwise to $\mathbb{1}\{x = 1\}$; the increments in (2) are bounded, so the limit passes through the expectation. The limiting increment for p is one exactly when the sources ahead of p in π leave some block uncovered and those up to and including p cover them all. Averaging over uniform π is the Shapley value of v_{all}, which is $\varphi^{\textsc{Completer}}$. □

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 12, in Section 8, 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 are the two ways of refusing to answer it in the middle.

5.5 Where the Shapley value comes in

Nothing so far has invoked a solution concept. Completer was defined as a Shapley value in Section 5.1 and Proposition 3 shows the reading model recovers it; the other two corners were derived from the reading model directly. We now record what their relation to the Shapley value is, because it is not the same for both.

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 $1/|P_b|$. Additivity of the Shapley value in the game closes the argument. The full proof is in Appendix A. □

So the two ends of the utility dial are the Shapley values of the two games, v_{lin} and v_{all}, and the dial interpolates between them. The order dial is different, and it is worth being careful about why. Every member of the family (2) is a random-order value in the sense of Weber (1988), and is therefore efficient and additive whatever π is. Only the uniform π is also symmetric. That matters here because the publishers of a block are interchangeable in v_{lin}: any symmetric value must split a block equally among them, so no Shapley value of v_{lin} can ever pay Provenance. Turning the order dial to chronological is precisely a refusal of symmetry, which is what it means to reward priority.

Provenance is nonetheless a Shapley value, of a third game. The three games differ only in what they say a group of publishers is worth. v_{all} asks whether the group can supply the whole answer. v_{lin} asks what share of the answer the group can supply. The third asks what share of the answer the group *broke*. Writing f_b for the publisher that filed block b first,

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

so a group is credited with a fact when one of its members was the first to report it, and carrying a fact 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 tell the reader the fact. Under v_{prov} the singleton $\{B\}$ is worth zero, because the fact 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.

Table 1: The two dials and the rules they generate.

	u linear *every fact counts*	u all-or-nothing *only a complete answer counts*
random order *veil of ignorance*	EQUAL (Prop. 1)	COMPLETER (Prop. 3)
chronological order *history rewarded*	PROVENANCE (Prop. 2)	— *pays the last gap-closer*

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

What (3) does is state a claim about desert as a claim about value: it says 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 order makes—there the later publisher earns nothing because it is always reached second, here because it never counted. The two are two ways of saying one 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 a matter of taste; the payments are the same either way.

5.6 The fourth corner rule

The two dials suggest 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 in Section 8.

5.7 A reservation about the completer rule

We record one concern about v_{all}. Because $v_{\text{all}} = \min_b v_b$ is one only when *every* block is covered, any publisher that uniquely supplies some cited block is a veto player: no coalition without it has any value at all. Publishers that supply only duplicated blocks are not excluded outright—one of them may still be the arrival that closes the last gap—but they are paid only for the orderings in which that happens, and the residual is small. Section 8.6 measures it: veto players collect 66.3% of the pool under the global index and 97.1% under the local one. The rule is therefore close to being a rule that pays for exclusivity alone, and under the local index it is almost exactly that rule, since 86.2% of cited blocks have a single publisher and nearly everyone is a veto player.

Veto status is also a knife-edge property of the index rather than of the world. One further outlet carrying an otherwise exclusive block removes a veto; one cluster that fails to merge two

wordings of the same fact creates one. Under EQUAL an error of that kind moves a payment 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. It is also what makes the rule easy to game, 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.

5.8 The third dial: which index the covering sets are read from

Everything above holds the covering sets P_b fixed. But P_b is a lookup, and what it returns depends on which index the marketplace asks. Against the *global index* of every article in the crawl, P_b is every outlet that carried the block. Against a *local index* built only from the pages a retriever surfaced for the query, P_b is only those of them that happen to appear in that window, and a block twenty outlets carried can look exclusive. Every rule in this section is defined over P_b, so every one of them is affected by which index it is read from. That choice is the third dial, and Section 6 shows it moves the payments more than either of the first two.

6 Why the global index is required

6.1 Two indexes, two counterfactuals

Every rule in Section 5 is defined over the covering sets P_b. Filling in 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?

The local index. P_b is restricted to the sub-index reachable from this 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?[3]

The distinction is invisible when facts are unique and decisive when they are duplicated. Our data are emphatically the latter case, and Section 8 shows the consequence: under the global index the rules of Section 5 separate, while under the local index they collapse onto each other. **The reason attribution rules matter at all is duplication, and duplication is only visible corpus-wide.**

[3]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.

Table 2: The three costs of running a content marketplace.

	Scales with	Index time (ours)	Inference time
C_1 *indexing*	per document, once	high	0
C_2 *answering*	per query	-52.3%	baseline
C_3 *attribution*	per query	exactly 0	high

C_1 decomposes a URL into a block tree and merges it into the index. C_2 builds the prompt and generates the answer; the index-time figure is measured in Section 8. C_3 determines who is paid.

6.2 Index time and inference time

A second, related distinction concerns *when* attribution is computed.

Inference-time attribution waits for a query and then works out which of the retrieved documents caused which parts of the generated answer. Three existing approaches do this: 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 does the work 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 nothing has to be inferred: the answer cites block identifiers, and settlement reads the covering sets off the index and divides the fee. This is the pipeline of Sections 3 and 4.

Timing and scope are different distinctions, but they are not independent. Working out attribution at inference time forces the use of the local index, because it is impractical to recalculate the global index for every query. By building a global index we could still use the local index for attribution; we show below, however, that the local index effectively collapses all payment rules and that the global index is required to separate them.

6.3 Cost: there are three bills, not one

Indexing is often considered to be expensive compared to inference. For a publisher content marketplace this is the wrong accounting, because a marketplace pays three separate bills.

Write D for documents, Q for queries, c_{gen} for the token cost of answering from raw article text, c_{dec} for the cost of decomposing one document, and c_{att} for the cost per query of determining attribution at inference time. Then

$$\underbrace{D\,c_{\text{dec}} + 0.48\,Q\,c_{\text{gen}}}_{\text{index time}} \qquad \text{versus} \qquad \underbrace{Q\,(c_{\text{gen}} + c_{\text{att}})}_{\text{inference time}}, \tag{4}$$

so the two paradigms cross at

$$Q^* \;=\; \frac{D\, c_{\text{dec}}}{0.52\, c_{\text{gen}} + c_{\text{att}}}. \tag{5}$$

Three observations follow.

The first is that the index need not be justified by attribution alone. The term $0.52\, c_{\text{gen}}$ sits in the denominator of (5) on its own. Even compared against a system that performs no attribution whatsoever, $c_{\text{att}} = 0$, the block representation still repays its indexing cost after $Q^* = D\, c_{\text{dec}}/(0.52\, c_{\text{gen}})$ queries, simply because answers generated from deduplicated blocks are cheaper to produce. We do not lean on this—the measurement comes from a single pipeline and would not survive a change of generation stack—but it does dispose of the objection that attribution is being bought at the price of efficiency.

The second is that inference-time attribution pays two per-query bills forever. Writing k for the number of documents retrieved for a query, perturbation methods require $O(k)$ to $O(k \log k)$ additional forward passes; Shapley over retrieved documents requires many more. Nematov et al. (2025) identify this as the binding constraint, noting that each utility evaluation involves an expensive model call with direct monetary and time costs. So c_{att} is large and grows with retrieval depth, and it is paid on every query for the life of the system, on top of the full-price C_2. Our C_1 is paid once per document.

The third is the one we think matters most: the cost argument and the counterfactual argument are the same argument. It is worth asking *why* our C_3 is zero. It is zero because the covering set is already materialized in the index, so settlement reduces to set arithmetic—and that is also what makes the *global* index affordable. Evaluating a global-index rule at inference time would require reasoning about every publisher who ever reported the fact, not merely the k that were retrieved, which is hopeless with a model in the loop. Inference-time methods are thus confined to the local index not by oversight but by budget. Cost is the reason the counterfactual is wrong.

Figure 9 plots the ledger.

6.4 Five further reasons index-time computation dominates for settlement

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 8 shows that

(a) Where each bill is paid

	scales with	index-time	inference-time
C_1 indexing	per document	high	≈ 0
C_2 answering	per query	-52%	baseline
C_3 attribution	per query	exactly 0	high

Index-time pays C_1 once per document.
Inference-time pays C_2 and C_3 on every query, forever.

(b) When the index repays itself

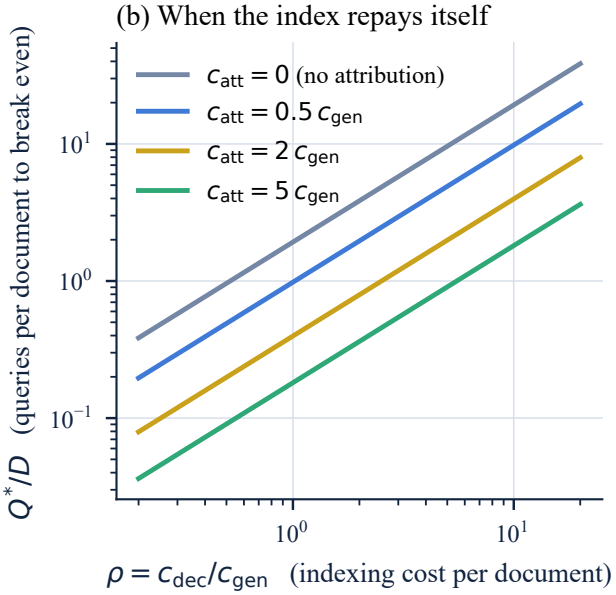


Figure 9: The three-cost ledger. Panel (a): where C_1, C_2 and C_3 are paid under each paradigm. Panel (b): the crossover Q^*/D against the indexing ratio $\rho = c_{\text{dec}}/c_{\text{gen}}$, for several per-query attribution costs $c_{\text{att}}/c_{\text{gen}}$.

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 10 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.

6.5 Counter-arguments to the global index

Three concessions are due. First, *extraction error*: index-time attribution inherits whatever mistakes the decomposition and clustering make, and inference-time methods have no such dependency; Section 10 reports our robustness checks. Second, *the index must be maintained*, and C_1 is a real and recurring bill: fresh content must be indexed before it can be attributed, which is a latency the inference-time approach does not have. Our defense is the ledger above rather than a denial—the cost is per document and amortizing—but an index serving few queries over a fast-churning corpus might be the wrong place to use our method. Third, *the two methods are complements*: inference-time attribution remains the right tool for verifying that a model actually used what it cites, and a mature marketplace would plausibly use inference-time methods to police the citation step and index-time methods to settle attribution.

7 Setting and data

7.1 The scenario

Our data come from a deployed prototype of a block-level marketplace, run over coverage of the Iran/US war from 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. The episode is well suited to the question: 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.

The scenario comprises 1,994 user queries, of which 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.

7.2 Premium classification

Throughout we report the share of value accruing to *premium* publishers—broadly, established outlets with substantial original newsgathering operations, as against aggregators, syndicators and low-cost republishers.

The 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, 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—and the tier was assigned before any attribution rule was computed, so it cannot have been fitted to the results.

The classification is not a normative judgment about quality. It is a proxy for the cost structure the marketplace is meant to affect: outlets that pay reporters are the ones for whom the difference between being paid for being first and being paid for being present is material. It is nonetheless the least mechanical construct in the paper, and we report in Appendix B how the headline moves when the two premium levels are separated and when the tier is replaced by the independent *type* coding, which identifies wire services, national news and broadcast outlets without reference to prestige.

Two caveats about the publisher table itself. 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 rule. Second, 1,126 of the 1,375 domains fall in the residual *other* type, which is a reminder that the long tail of the crawl is mostly not recognizable news organizations at all.

7.3 Publication timestamps

The PROVENANCE rule requires knowing who published first, so the quality of publication timestamps determines how much weight the corresponding results can bear. We are more exposed here than we would like.

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 sit at exactly 00:00:00 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.

Two points must be made plainly. First, the asymmetry runs *against* premium outlets: they are the group whose exact filing times are least often recorded, so imputation is not a neutral operation. Second, because PROVENANCE is decided by an ordering, and because a third of the relevant timestamps are imputed, any single convention for handling them is a modeling choice that could drive the result.

We therefore do not adopt one. Section 8 reports the PROVENANCE rule under four conventions—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—and asks the reader to judge the result by its stability across them rather than by any one number. Ties are always split equally.

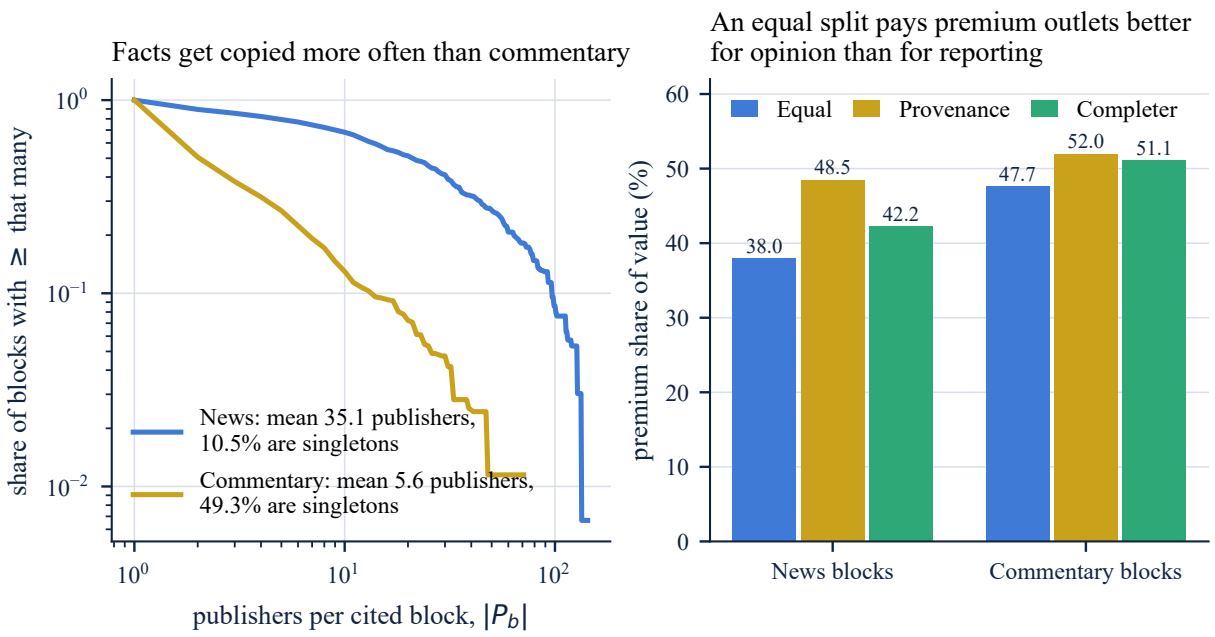


Figure 10: The duplication gap. Left: survival function of publishers per cited block in the global index, news against commentary. Right: the same gap in payment terms—an equal split pays premium outlets nearly ten points more for opinion than for reporting.

8 Results

8.1 The index is full of duplication, and that is the point

Begin with the structure of the covering sets, since everything else follows from it. 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 not an edge case; it is the normal condition of a cited fact.

It is also very unevenly distributed across the two block kinds. News blocks are singletons 10.5% of the time, commentary blocks 49.3% of the time: facts get copied more often than the commentary written about them. Figure 10 reports the full distributions. This is the microfoundation for everything that follows: a payment rule is a position on what to do about duplication, so it can only matter to the extent that duplication occurs.

8.2 Provenance favors premium publishers

Table 3 reports the premium share of value under the two rules that the order dial connects, computed against the global index, under each of the four timestamp conventions of Section 7.3.

Rewarding first publication moves between five and eight percentage points of total value toward premium publishers, and—this is the part that matters given Section 7.3—the sign and rough magnitude survive every convention we tried, including the most conservative one, which discards imputed times altogether. The mechanism is priority: premium outlets are

Table 3: 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.

Table 4: 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 3.

disproportionately the ones who 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 about what they are owed.

We resist a stronger claim. The spread is real but moderate, and it is not a license to describe PROVENANCE as transformative. What Table 3 establishes is that the choice of rule is *consequential*—enough to be worth making deliberately—not that one rule is correct.

8.3 The index scope dominates the rule

Now change the third dial and hold everything else fixed. Table 4 recomputes the same quantities against the local index.

Inside the local index, 86.2% of cited blocks have exactly one publisher and mean coverage is 1.23. There is almost nothing for a rule to disagree about: every rule and every timestamp convention lands within a single percentage point of every other. Figure 11 shows the two coverage distributions side by side.

We regard this as the paper's most important empirical result, and it is stronger than the usual framing of scope as a robustness check. It is not that the local index gives a somewhat different answer. It is that inside the local index **no attribution policy is expressible at all**. Whatever normative position a marketplace takes—reward availability, reward originality, reward indispensability—it will pay out the same money. Since an inference-time method can only ever observe the retrieved set, this is a statement about what such methods can do in principle, not

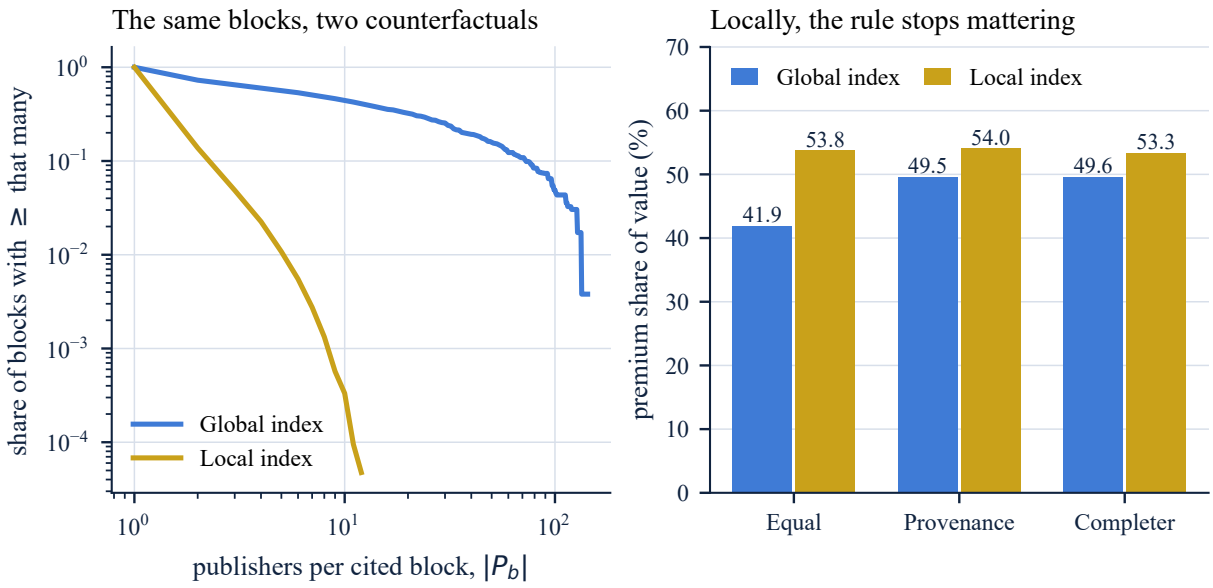


Figure 11: Global against local index. Left: coverage distributions under the two scopes. Right: the premium share under each rule. Locally, the rules collapse onto each other—there is nothing left for a payment rule to disagree about.

about how well they do it.

8.4 The utility dial

Hold the order random and turn the second dial. Let the buyer's utility be $u(x) = x^\alpha$ in the fraction x of the answer's blocks covered, so that $\alpha = 1$ is a buyer who values every fact separately and $\alpha \to \infty$ is a buyer who values only a complete answer. Payments are the expected marginal contributions of Section 5, estimated by sampling 200 publisher orderings per query.

Table 5 reports the path, and Figure 12 draws it. Three things are worth drawing out.

The first is a check rather than a result. The sampled $\alpha = 1$ row reproduces the closed-form EQUAL share to 0.01 percentage points, and the whole publisher payout vector to 0.57 points, which is what Proposition 4 requires: the equal per-block split is not merely similar to a Shapley value, it *is* the Shapley value of the linear coverage game. That the sampler rediscovers it without being told is the strongest evidence we can offer that the unification is real and not a coincidence of aggregates.

The second is that the premium share rises monotonically in α, from 41.87% to 49.56%, and that 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. The dial has an economic reading and the data respect it.

The third is the one we did not expect. The utility dial and the order dial arrive at nearly the same aggregate from opposite directions: COMPLETER pays premium publishers 49.56%

Table 5: 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.

and PROVENANCE pays them 49.55%, a difference of less than a tenth of a point. These are economically opposite rules—one rewards being irreplaceable, the other rewards being early—and they agree on the headline. They do not, however, agree on *who gets paid*: the total variation between their publisher-level payout vectors is 33.7 percentage points, larger than the distance from either of them to EQUAL. The aggregate agreement is a coincidence of composition, and a reader who compared only the headline numbers would conclude the two rules were interchangeable when at the publisher level they are further apart than any other pair we compute. We flag this as a caution about evaluating allocation rules by their aggregates.

The fourth corner, and the two random orders

Two further cells of the family are worth reporting briefly. 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.

Separately, Section 5.3 distinguished two ways of randomizing: over *articles* and over *publishers*. Only the second gives the EQUAL rule exactly. 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 volume subsidy rather than a value judgment. We use the publisher-randomization throughout, and note that a marketplace which reached for the "obvious" article-level randomization would be paying for output rather than for content.

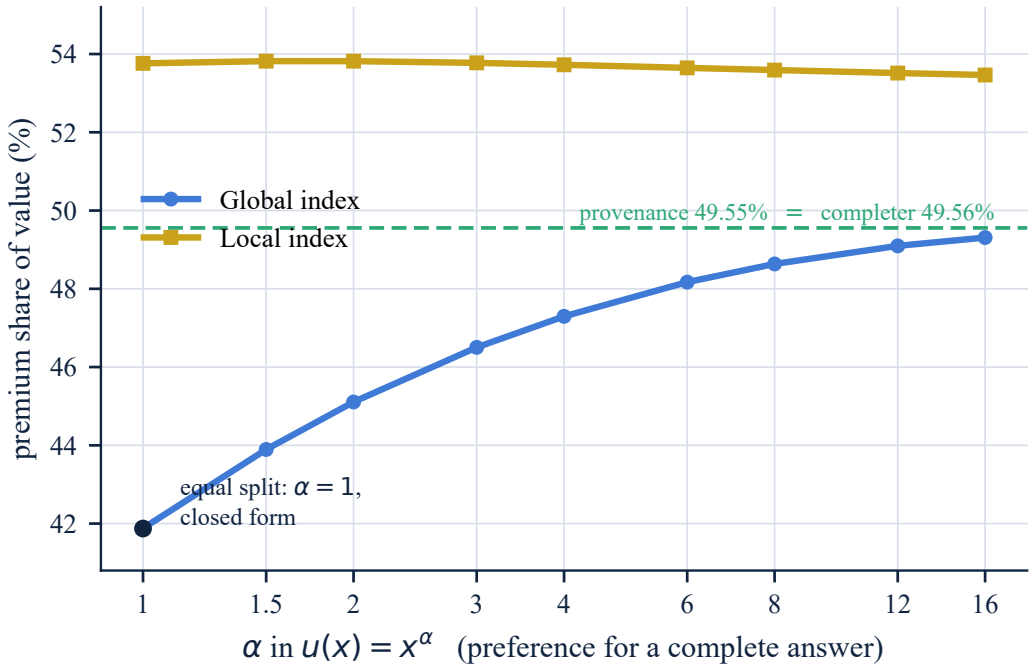


Figure 12: The utility dial. Premium share of value against α in $u(x) = x^{\alpha}$, random order, under both index scopes. Provenance and the completer land within a hundredth of a point of one another and are drawn as a single reference line—from opposite economic directions.

8.5 The news and commentary split

The premium share differs sharply by block kind, and the difference tracks the duplication gap of Section 8.1 exactly.

A cited news block is carried by 35.1 publishers on average; a cited commentary block by 5.6. The consequence for payment is large and, we think, uncomfortable. 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 is the sharpest statement of the paper's economic point that we can make from the data. **An equal split does not merely fail to reward original reporting; it actively prices reporting below commentary, because the property that makes reporting valuable—that others find it worth copying—is the same property that dilutes its payment.** Whether a marketplace should correct this is a question we leave open. That the default rule has this incidence is not something a designer should discover after the fact.

Table 6: 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 7: 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.

8.6 Concentration

Payment rules that differ by a few points in the premium aggregate can differ a great deal in how widely they spread money, and the rules here do.

The top fifty publishers take 50.1% of the pool under EQUAL and 68.5% under PROVENANCE; the Gini 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—it is a substantial concentration of the pool, and the premium share understates it, because the movement is toward a small number of large originators rather than toward the premium tier as a class.

The last row is the one we would put in front of a market designer. 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 and more troubling reason, which Section 5.7 anticipated. Because v_{all} is one only when every cited block is covered, a publisher that uniquely supplies some cited block 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

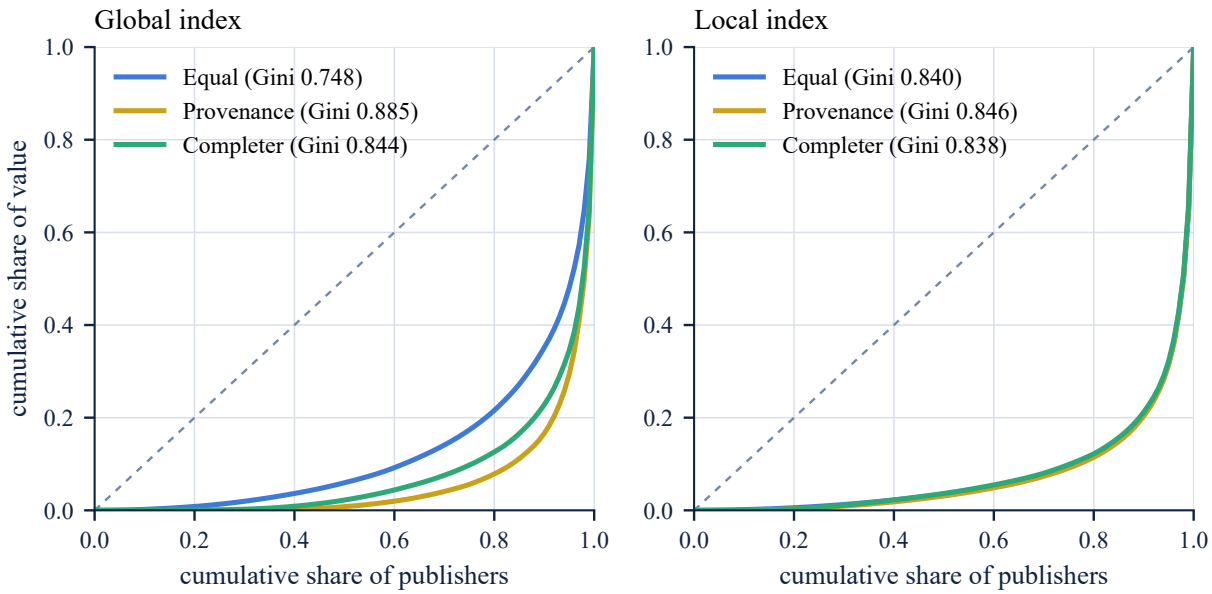


Figure 13: Concentration. Lorenz curves of value across publishers under each rule, on a fixed base of contributing publishers. Under the local index the three curves are indistinguishable.

local index turns it into one almost exactly. This is our main reason for treating it as the limiting case of a dial rather than as a candidate policy.

Figure 13 reports the full distribution of value across publishers under each rule, with the premium tier overlaid, and Figure 14 reports the veto-player share alongside the top-k curves.

8.7 Publisher-level results

Everything so far has been an aggregate. Table 8 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 are simply everywhere. 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 what the order dial does. 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 8.1: 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 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.

One row should not be passed over. Specialized information site #1 is 95th under Equal

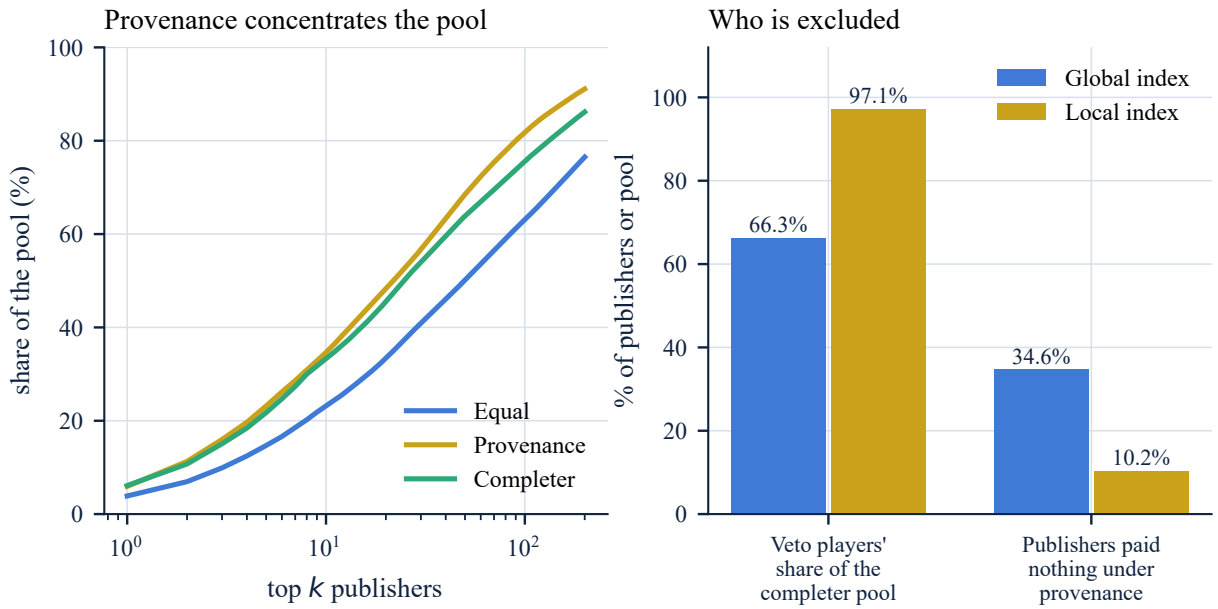


Figure 14: Who gets paid. Left: top-k share of the pool under each rule, global index. Right: the share of the completer pool collected by veto players, and the share of contributing publishers paid nothing at all under provenance.

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 it broke those facts. We believe it carries filing times that sort early, and Section 7.3 explains why that is not a remote possibility. This is the fragility of a first-mover rule made concrete: it is only as good as the timestamps it is decided by, and the failure mode is not a small distortion of the aggregate but a large payment to the wrong party. Note finally that these ten publishers collect 28.8% of the pool under PROVENANCE while the top ten *under* PROVENANCE collect 34.6%: the rule does not merely rescale the ranking, it rewrites it.

9 Incentives and market design

A payment rule is not only a division of a fixed pie; it is a signal about what to produce. This section takes the rules of Section 5 as candidate policies and asks what each one would do to publishers if it were adopted.

9.1 What each rule funds

EQUAL pays for availability. A publisher is paid for having the fact, regardless of where it got it, so the rule funds distribution and is indifferent between the outlet that sent a reporter and the outlet that rewrote the wire copy twenty minutes later. Its virtue is that it is robust: it needs no timestamps, cannot be gamed by manipulating filing times, and is trivially auditable.

Table 8: 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.

PROVENANCE pays for originality. It funds newsgathering directly, which is the stated purpose of most proposals in this space and aligns with the diagnosis in Cagé et al. (2020) that original reporting is under-rewarded precisely because copies arrive so quickly. Its costs are the subject of the next subsection.

The COMPLETER rule pays for indispensability—for holding something nobody else has. It has an appealing economic logic, since it prices exactly the counterfactual loss the buyer would suffer, but it is also the most concentrating of the three, and we regard its distributional properties as its principal weakness (see Deng and Papadimitriou, 1994).

9.2 Racing

The clearest hazard of PROVENANCE is that it makes speed the object of competition. This is a classic patent-race structure (Loury, 1979; Lee and Wilde, 1980; Fudenberg et al., 1983): when the entire value of a fact accrues to whoever files first, effort is diverted from verification toward speed, and the resulting equilibrium can be socially wasteful even when each firm's behavior is privately optimal. Hill and Stein (2025) document the quality consequences of exactly this dynamic in scientific publishing.

The size of the prize matters for how seriously to take the concern, and Section 8.2 bounds it:

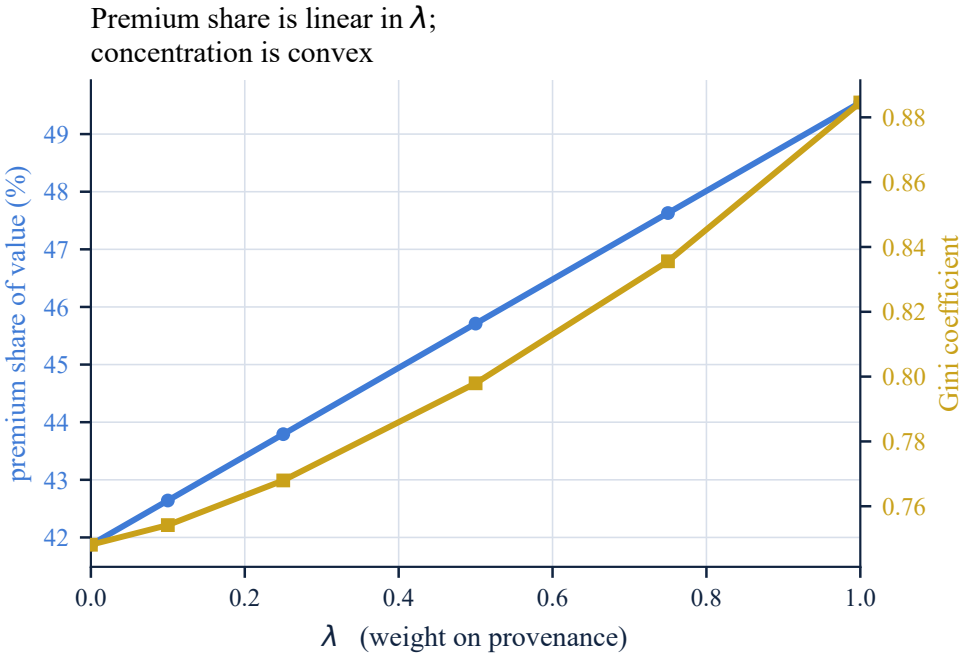


Figure 15: Partial provenance. Premium share is exactly linear in the provenance weight λ, while concentration is convex. Small λ therefore buys redistribution cheaply in concentration terms, and keeps every contributing publisher strictly paid.

five to eight percentage points of aggregate share. That is large enough to change behavior and small enough that partial measures are credible. Three suggest themselves: a verification lag, so that provenance credit vests only after some interval; a clawback for content subsequently shown to be false; and a partial provenance weight λ that interpolates between EQUAL and PROVENANCE.

The third of these is worth taking seriously, because the interpolation behaves unusually well. Since both rules distribute exactly one unit per query, so does the blend $\varphi^{\lambda} = (1-\lambda)\varphi^{\text{EQUAL}} + \lambda\varphi^{\text{PROVENANCE}}$, and the premium share is therefore *exactly* linear in λ: it traverses 41.87% to 49.55% in a straight line, with a maximum deviation from linearity of order 10^{-14} percentage points. A designer who wants to reward originality by a specified amount can simply read λ off that line. At $\lambda = \frac{1}{4}$ the premium share is 43.79%, at $\lambda = \frac{1}{2}$ it is 45.71%.

Concentration, by contrast, is convex in λ: the Gini moves from 0.748 to 0.768 at $\lambda = \frac{1}{4}$ and 0.798 at $\lambda = \frac{1}{2}$, against 0.885 at the endpoint. The first quarter of the way toward PROVENANCE therefore buys about a quarter of the redistribution toward premium publishers but only about a seventh of the increase in concentration, and—more to the point for the racing concern—it leaves every publisher with a strictly positive payment, whereas full PROVENANCE pays 34.6% of contributing publishers nothing. Partial provenance is not a compromise between two good rules. It is strictly better than either on the participation margin, and it dulls the race prize proportionately.

9.3 Timestamp manipulation

If filing time determines payment, filing time becomes strategic. Publishers control their own reported publication times, and Section 7.3 shows those times are already of uneven quality. A marketplace adopting PROVENANCE would need timestamps attested by the platform at crawl time rather than self-reported—which is feasible, but is an infrastructural requirement rather than a detail, and it is one more reason the settlement layer has to be an index rather than an inference-time computation.

9.4 Duplication attacks, and why deduplication is the defense

Any rule that pays per retrieved chunk invites the obvious attack: republish the same content across many properties so that more chunks are retrieved. Cross-publisher clustering neutralizes this by construction, since a block is credited once regardless of how many documents carry it, and the credit is divided among publishers rather than multiplied by them. This is a structural advantage of the block index that has nothing to do with which rule is chosen, and it applies with equal force under all three.

9.5 The music analogy, and its limits

Streaming faced the same problem—divide a subscription pool among rights holders—and the two rules it actually uses, pro-rata and user-centric, map onto our order dial closely enough to be instructive (Datta et al., 2018; Aguiar and Waldfogel, 2021). The analogy has one important limit. A song is a single work with a single owner; a fact is non-rival and frequently held by twenty parties at once (Jones and Tonetti, 2020). The allocation problem in news is therefore not merely a division problem but a *duplication* problem, which is why the covering set, and hence the index, is doing work that has no counterpart in music.

10 Limitations

We state the things that most constrain what can be concluded.

The 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, that the index scope dominates the rule—to generalize.

Publication timestamps are substantially imputed. As Section 7.3 sets out, only 33.3% of premium and 58.4% of non-premium grounded URLs carry an exact time. Our defense is to report four conventions rather than to select one, and the result is stable across them; but a

rule decided by an ordering, applied to data a third of which is imputed, deserves the reader's skepticism and we prefer to invite it explicitly.

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 often enough. 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 and we cannot test it here.

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%, which is well beyond any plausible error rate for this pipeline. The headline is untouched: the PROVENANCE–EQUAL spread stays between +7.62 and +8.12 percentage points against a baseline of 7.68. We report this because the worry is a reasonable one, not because it turned out to matter.

Finally, there is what we do not measure at all. 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 Section 9 is a discussion of incentives rather than an equilibrium analysis. And we 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.

11 Conclusion

The click is disappearing as a unit of account, and something has to replace it. The proposals now circulating agree that publishers should be paid for content the model actually uses, and

disagree—mostly implicitly—about what "uses" means.

We have argued three things. First, that the apparent disagreement about how to pay is smaller than it looks. An equal split, a first-mover rule and a rule that pays for indispensability all emerge from a simple model of news reading, under two dials—a reading order and a utility curvature—each with a plain economic reading, and each rule can be mapped to a Shapley value. The equal split is exactly the Shapley value of a linear coverage game; the Shapley value used in the applied literature is the same construction with an all-or-nothing buyer. We do not offer this as a theoretical advance. We offer it because it is what allows attribution to be *measured* rather than merely computed: comparing formulas tells you they differ, whereas locating them in a space tells you along which dimension, and converts the design problem into two questions a market designer can be asked directly—should history be rewarded, and how much does the buyer need the answer to be complete?

Second, that the choice is worth money but not a fortune. Moving from an equal split to a first-mover rule shifts five to eight percentage points of value toward premium publishers, robustly across four ways of handling incomplete timestamps. That is enough to change what gets funded, and enough to start a race worth worrying about, without being the transformation that advocacy on either side tends to imply.

Third—and this is the finding we would most like to survive—that a third dial dominates the other two. Whether a publisher's contribution is judged against the *global index* of everything published or against the *local index* of sources retrieved for one query decides whether any of the rest matters. 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 pays the same. A marketplace that settles against the local index has not chosen a lenient attribution policy; it has chosen not to have one.

That is what makes the architectural question an economic one. 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, and it costs nothing per query. Whatever rule a marketplace ultimately adopts, it will need an index to adopt it against.

Finally, a note on what a settlement layer requires that a metering scheme does not. Publishers cannot verify a saliency map, but they can be shown their rows in an index: these are your blocks, this is who else carried them, this is when you filed. Contracts are written on verifiable states. If the goal is a market rather than a gesture, the auditability of the measurement is not a feature of the design—it is the design.

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.

Xiaotie Deng and Christos H. Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of Operations Research*, 19(2):257–266, 1994. doi: 10.1287/moor.19.2.257.

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.

Drew Fudenberg, Richard Gilbert, Joseph Stiglitz, and Jean Tirole. Preemption, leapfrogging and competition in patent races. *European Economic Review*, 22(1):3–31, 1983. doi: 10.1016/0014-2921(83)90087-9.

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.

Ryan Hill and Carolyn Stein. Race to the bottom: Competition and quality in science. *Quarterly Journal of Economics*, 2025. doi: 10.1093/qje/qjaf010.

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.

Tom Lee and Louis L. Wilde. Market structure and innovation: A reformulation. *Quarterly Journal of Economics*, 94(2):429–436, 1980. doi: 10.2307/1884551.

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.

Glenn C. Loury. Market structure and innovation. *Quarterly Journal of Economics*, 93(3): 395–410, 1979. doi: 10.2307/1883165.

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.

Robert J. Weber. Probabilistic values for games. In Alvin E. Roth, editor, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, pages 101–119. Cambridge University Press, 1988. doi: 10.1017/CBO9780511528446.008.

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 9: 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 10: 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 7.3 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 8.2 exactly, which is the check that it is measuring what we intend.

Table 10 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 8.4 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 16 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 8.4. 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 8.4, 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 Implementation notes

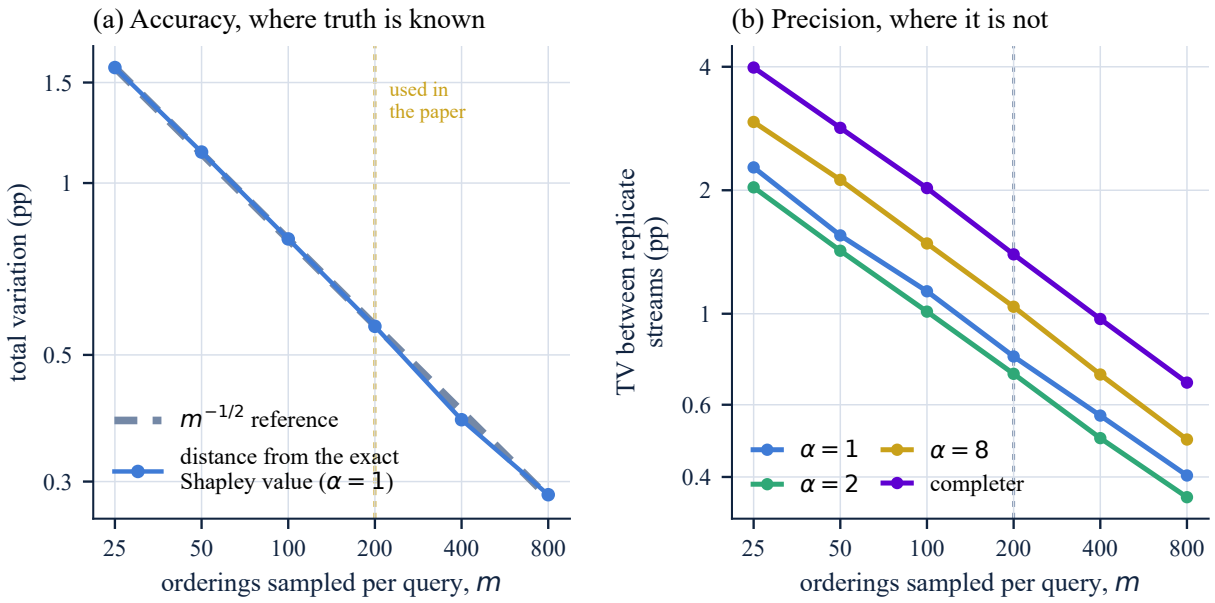


Figure 16: Monte Carlo convergence. Panel (a): distance from the exact Shapley value at $\alpha = 1$, against a $m^{-1/2}$ reference the data are indistinguishable from. Panel (b): mean pairwise distance between three independent streams, for rules with no closed form. The dashed vertical line marks the $m = 200$ used throughout the paper.