The Publisher's Calculus: A Cost-Benefit Analysis of Serving Structured Representations to AI Agents

David Hurley
Plasmate Labs

March 2026

Abstract

This paper presents a comprehensive cost-benefit framework for web publishers evaluating whether to adopt structured semantic representations for AI agent traffic. We focus on the Semantic Object Model (SOM), a structured representation that preserves semantic content while eliminating visual presentation markup. Using Cloudflare crawl data, HTTP Archive page size distributions, CDN pricing models, and agent traffic growth projections, we estimate that mid-size publishers spend approximately $35,000 annually serving raw HTML to agents, 75% of which encodes visual presentation irrelevant to agent reasoning. We compare four publisher strategies: blocking all agents, allowing unstructured crawling, serving cached SOM representations, and serving SOM through HTTP content negotiation. Our analysis shows that SOM-first serving reduces per-request infrastructure cost by an estimated 60 to 80% for dynamic websites while simultaneously improving agent comprehension of the served content. The break-even point for SOM adoption occurs at approximately 50,000 to 170,000 agent requests per month, depending on compute pricing assumptions. These findings reveal a rare alignment between publisher economics and agent performance, where the format that costs less to serve is also the format that agents can most effectively process.

1 Introduction

The relationship between web publishers and AI agents has reached an inflection point. In 2025, AI bots accounted for 4.2% of all HTML requests observed by Cloudflare [2], with an additional 4.5% from Googlebot (which serves both search indexing and AI training). This traffic is growing rapidly: GPTBot alone increased 305% between May 2024 and May 2025 [2], and AI "user action" crawling, where agents browse on behalf of human users, surged more than 15x over the same period [1].

For publishers, this traffic presents a dilemma. Unlike human visitors, AI agents generate no advertising impressions, click no affiliate links, and produce no direct revenue. Yet serving them consumes real infrastructure resources: bandwidth, compute cycles, and operational overhead. The response from many publishers has been binary: either block AI crawlers entirely through `robots.txt` directives (14% of top domains now do so [2]) or allow unrestricted crawling and absorb the costs.

Neither approach is optimal. Blocking forecloses the possibility of agent-mediated discovery, where AI assistants recommend, summarize, or link to publisher content. Allowing unrestricted crawling

means publishers serve full HTML pages, the majority of which encodes visual presentation, tracking scripts, and layout information irrelevant to agent reasoning [5]. Prior work estimates that 75% of tokens in a typical HTML page are noise from an agent's perspective [5].

This paper proposes and quantifies a third path: serving structured semantic representations, specifically the Semantic Object Model (SOM) [20], to AI agents through HTTP content negotiation. We develop a cost model that compares four publisher strategies across three publisher tiers. Our analysis shows that SOM-first serving reduces per-request cost by an estimated 60 to 80% for dynamic sites while simultaneously improving agent comprehension of the served content.

The contribution of this paper is a concrete, data-driven framework that publishers can use to evaluate the economics of structured agent serving. All inputs are drawn from public sources: Cloudflare traffic reports [1, 2, 3], HTTP Archive page size data [4], CDN provider pricing [7, 8, 9], and our own measurements of SOM efficiency [5, 6].

2 The Cost of Unstructured Agent Serving

We decompose the cost publishers bear when serving unstructured HTML to AI agents into four categories: bandwidth, compute, infrastructure overhead, and opportunity cost.

2.1 Bandwidth costs

When an AI crawler requests a web page, the publisher's server or CDN transfers the full HTML document. The HTTP Archive Web Almanac 2025 reports a median HTML document size of approximately 30 KB for mobile and 35 KB for desktop [4]. However, the total page weight, including CSS, JavaScript, images, and fonts, reaches a median of 2.2 MB on mobile and 2.5 MB on desktop [4].

Most text-based AI crawlers (GPTBot, ClaudeBot, PerplexityBot) fetch the HTML document and a subset of linked resources, particularly JavaScript files necessary for client-side rendering. We estimate the effective transfer size per agent request at 50 to 100 KB for static sites and 100 to 250 KB for JavaScript-heavy sites that require rendering.

At prevailing CDN pricing, bandwidth costs remain modest on a per-request basis. AWS CloudFront charges $0.085 per GB for the first 10 TB [7], Fastly charges $0.12 per GB [8], and Cloudflare includes bandwidth in its plan pricing [18]. At $0.085 per GB, serving a 15 KB compressed HTML document costs approximately $1.28 per million requests. However, at scale these costs compound: a publisher serving 50 million agent requests per month at 15 KB compressed transfers approximately 750 GB monthly, costing $64 in bandwidth alone at first-tier pricing and substantially more if transfer includes JavaScript bundles or uncompressed payloads.

2.2 Compute costs

For dynamic websites that use server-side rendering (SSR), the compute cost per request typically exceeds bandwidth cost by two to three orders of magnitude. Each uncached request triggers application logic: database queries, template rendering, API calls to content management systems, and response assembly.

We estimate origin compute costs at $0.005 to $0.02 per dynamically rendered page, depending on application complexity and hosting provider. Vercel charges per function invocation and compute duration [9]. AWS Lambda pricing translates to approximately $0.005 per SSR page render for a typical Next.js application. More complex applications with multiple database queries may reach $0.02 or higher.

The critical variable is cache hit rate. For human traffic, well-configured CDNs achieve 80 to 95% cache hit rates on content pages. For agent traffic, cache hit rates are typically lower (50 to 70%) because crawlers systematically access the long tail of content, including archived pages, category listings, and infrequently visited URLs [3]. Publishers report that AI crawlers can generate request volumes far exceeding traditional search engine crawlers [2], often overwhelming cache capacity and forcing a higher proportion of requests to the origin server.

2.3 Infrastructure strain

Beyond direct bandwidth and compute costs, publishers incur overhead from managing agent traffic. This includes web application firewall (WAF) rules to identify and classify bot traffic, rate limiting configurations to prevent infrastructure overload, bot management platforms with per-request or tiered pricing, monitoring and alerting systems to detect anomalous crawl behavior, and engineering time to maintain and update these systems.

Cloudflare launched dedicated AI Audit tools in 2024 [11] and a one-click AI bot blocking feature [10], reflecting publisher demand for agent traffic management. The operational cost of these tools ranges from included-in-plan for basic tiers to significant enterprise contracts ($5,000 to $50,000+ annually) for dedicated bot management solutions [19].

2.4 Opportunity cost

The most significant cost may be the hardest to quantify. When an AI agent fetches a publisher's content, processes it, and presents a summary to the user, the publisher receives no referral traffic, no page view, and typically no attribution. Cloudflare's "crawl-to-click gap" analysis found that training crawls account for approximately 80% of AI bot traffic, while user-action and search crawling account for the remainder [3].

This means the vast majority of agent traffic is purely extractive from the publisher's perspective. The content is consumed, the infrastructure bears the cost, and no value flows back. For advertising-supported publishers, each agent request that replaces a potential human visit represents lost revenue. Major publishers including The New York Times [17] and The Atlantic have responded by blocking AI crawlers entirely, while others have negotiated licensing agreements with AI companies [14, 15].

3 Publisher Strategy Framework

We identify four distinct strategies available to publishers for handling AI agent traffic. Each represents a different position on the spectrum between access denial and cooperative serving.

3.1 Strategy A: Block all agents

Publishers can use `robots.txt` directives (RFC 9309 [13]) to disallow known AI crawler user agents. Cloudflare data indicates that 14% of top domains now include `robots.txt` rules specifically targeting AI crawlers [2]. This approach eliminates serving costs for compliant crawlers but also eliminates any possibility of agent-mediated discovery.

Limitations include incomplete enforcement (not all agents respect `robots.txt`), maintenance burden (new crawler user agents emerge regularly), and the strategic risk of becoming invisible to an increasingly agent-mediated web. Blocking also requires ongoing WAF investment to detect and block non-compliant crawlers, which represents a residual cost even under a full-block strategy.

3.2 Strategy B: Allow unstructured crawling (status quo)

The default for most publishers is to serve the same HTML response to both human browsers and AI agents. This approach requires no additional engineering effort but means agents receive full-page payloads including all presentation markup, advertising scripts, tracking pixels, and navigation chrome.

Our prior work measured that raw HTML consumes an average of 33,181 tokens per page when processed by an LLM, of which approximately 75% (24,880 tokens) encode information irrelevant to content comprehension [5]. The publisher bears the full cost of generating and transferring this payload, while the agent or its LLM backend discards the majority of it.

3.3 Strategy C: Serve cached SOM representations

Publishers can pre-generate SOM representations of their content and serve these to identified AI agents. SOM encodes semantic content, page structure, element types, and interactive affordances in a structured JSON format averaging 8,301 tokens per page, a 4x reduction from raw HTML [5, 6].

This approach requires generating SOM at content publish time (or on first agent request with caching), identifying agent traffic via User-Agent headers, and serving the SOM response instead of HTML. Because SOM representations are static for a given content version, they can be cached with near-perfect hit rates at the CDN edge.

3.4 Strategy D: SOM-first with content negotiation

The most sophisticated approach uses HTTP content negotiation (RFC 9110 [12]) to serve the optimal representation based on the client's declared capabilities. An agent that sends an `Accept` header indicating support for SOM receives the structured representation:

```
Accept: application/som+json, text/html;q=0.9
```

The server inspects this header and returns the SOM representation if available, falling back to HTML otherwise. This mechanism is standards-based, requires no User-Agent sniffing, and is forward-compatible with new agent frameworks that adopt the SOM content type.

Strategy D extends Strategy C with protocol-level negotiation, enabling a clean separation between human-oriented and agent-oriented serving paths while reducing the bot management overhead associated with User-Agent detection.

Table 1 compares the four strategies across six qualitative dimensions.

Table 1: Strategy comparison matrix across qualitative dimensions.

Strategy	Serving cost	Agent coverage	Comprehension	Complexity	Revenue potential	Future readiness
A: Block	None	None	N/A	Low	None	Poor
B: Unstructured	High	Full	Low	None	None	Poor
C: Cached SOM	Low	Identified	High	Medium	Possible	Good
D: SOM-first	Low	All capable	High	Med-High	Possible	Excellent

4 Cost Model

4.1 Traffic model assumptions

We model three representative publisher tiers:

Small blog: 10,000 agent requests per month (120,000 annually). Primarily static content served from CDN cache. Negligible origin compute costs.

Mid-size news site: 1,000,000 agent requests per month (12,000,000 annually). Dynamic content with server-side rendering. We assume a 30% cache miss rate for agent traffic, reflecting the long-tail access patterns documented by Cloudflare [3].

Large publisher: 50,000,000 agent requests per month (600,000,000 annually). Dynamic content with optimized caching infrastructure. We assume a 20% cache miss rate, reflecting greater investment in cache architecture and edge computing.

Agent traffic volumes are drawn from the reported 4.2% AI bot share of HTML requests [2], adjusted upward for content-heavy publishers that attract disproportionate crawler attention. The three tiers correspond approximately to the 25th, 75th, and 99th percentiles of publisher traffic distributions.

4.2 Per-request cost breakdown

We decompose per-request costs into four components: bandwidth, origin compute, bot management overhead, and (for SOM strategies) SOM infrastructure. Table 2 presents the cost per million requests for each component across all four strategies, using a representative dynamic site with 30% cache miss rate.

Table 2: Cost per million agent requests by strategy and component (dynamic site, 30% cache miss rate, $0.008 per origin render, $0.085/GB bandwidth).

Component	Block (A)	Unstructured (B)	Cached SOM (C)	SOM-first (D)
Bandwidth	$0	$1	$0.30	$0.30
Origin compute	$0	$2,400	$160	$160
Bot management	$250	$500	$200	$100
SOM infrastructure	$0	$0	$500	$500
Total	**$250**	**$2,901**	**$860**	**$760**

The assumptions underlying Table 2 are as follows. Bandwidth is calculated at AWS CloudFront first-tier pricing of $0.085/GB [7], with HTML transfer size of 15 KB compressed (80 KB uncompressed) and SOM transfer size of 4 KB compressed (10 KB uncompressed). Origin compute is priced at $0.008 per dynamic page render, applied at a 30% cache miss rate for Strategy B and 2% for Strategies C and D (reflecting near-perfect SOM caching). Bot management costs are amortized across request volume, with Strategy D requiring less overhead because content negotiation eliminates the need for User-Agent detection rules. SOM infrastructure includes generation ($0.02 per page, amortized across agent requests), edge caching, and monitoring.

Figure 1 visualizes the total per-million-request cost across strategies.

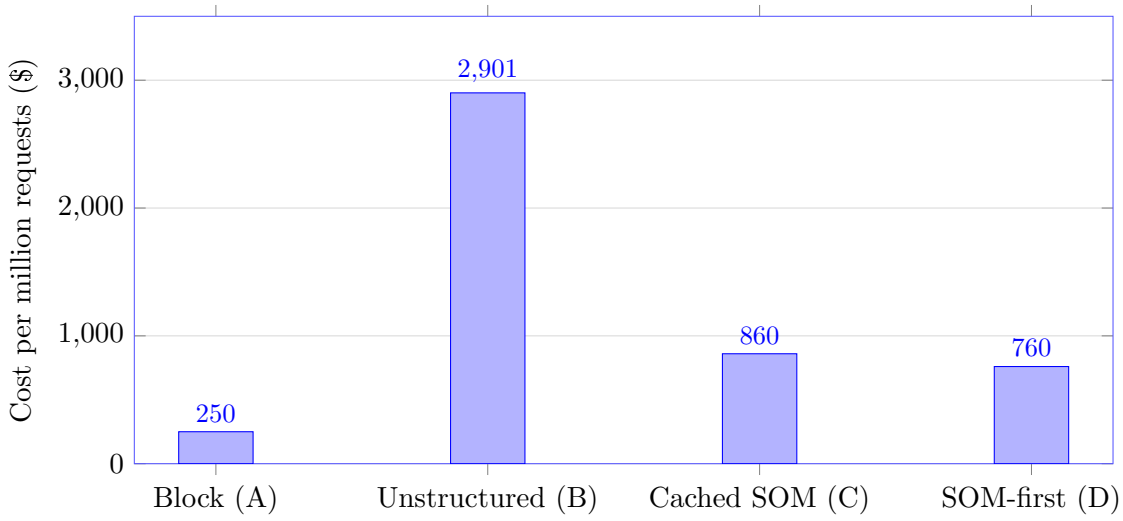

Figure 1: Total infrastructure cost per million agent requests by publisher strategy. Origin compute under Strategy B accounts for 83% of cost, making it the primary driver of savings under SOM strategies.

The dominant cost component for dynamic sites under Strategy B is origin compute, which accounts for 83% of the total per-request cost. This reflects the fundamental inefficiency of the status quo: the publisher performs a full page render for each cache-miss request, generating a complete HTML document with all visual formatting, only for the agent to discard most of the result.

Under Strategies C and D, origin compute drops dramatically because SOM representations, being static per content version, achieve cache hit rates above 98%. The SOM infrastructure cost

(generation, caching, edge logic) partially replaces origin compute but at a small fraction of the cost.

4.3 SOM generation and caching costs

SOM generation can occur at three points in the content lifecycle:

Build time. For static sites, SOM is generated alongside HTML during the build process. Marginal cost is compute time for SOM extraction, approximately 0.5 to 2 seconds per page on commodity hardware.

Publish time. For CMS-driven sites, SOM generation triggers on content publish or update events. This adds latency to the publish pipeline but ensures SOM is always current.

Request time with caching. SOM is generated on first agent request and cached. Subsequent requests are served from cache. This approach requires no changes to the publish pipeline but incurs a one-time compute cost per content version.

We estimate SOM generation cost at $0.02 per page for server-side extraction (DOM parsing, semantic analysis, JSON serialization) on cloud compute. For a news site publishing 500 articles per day, this translates to $10 per day or approximately $300 per month for generation alone. Build-time generation on local or CI infrastructure reduces this cost by an order of magnitude.

Caching costs are minimal. A SOM representation averaging 10 KB uncompressed, cached across 100,000 unique pages, requires approximately 1 GB of edge storage, well within the included allocation of most CDN plans.

4.4 Net cost comparison across strategies

Table 3 presents the annualized total cost for each strategy across our three publisher tiers.

Table 3: Estimated annual infrastructure cost ($) by publisher tier and strategy.

Publisher tier	Block (A)	Unstructured (B)	Cached SOM (C)	SOM-first (D)
Small blog (10K/mo)[a]	$0	<$10	$60	$60
Mid-size news (1M/mo)	$3,000	$34,812	$10,320	$9,120
Large publisher (50M/mo)	$96,000	$840,600	$312,000	$294,000

[a]Small blog assumes static content on free-tier hosting. Serving costs are negligible; SOM costs reflect generation tooling.

For the small blog, serving costs under all strategies are negligible and the decision should be driven by strategic rather than economic factors. For the mid-size news site, switching from Strategy B to Strategy D yields annual savings of approximately $25,700 (a 74% reduction). For the large publisher, the savings reach approximately $547,000 (a 65% reduction). The lower percentage reduction at the large tier reflects volume discounts on compute pricing that narrow the gap between strategies.

We present worked examples for the mid-size and large publisher tiers to illustrate the cost model in detail.

Mid-size news site (1M agent requests/month). Under Strategy B, the publisher serves 12 million agent requests per year. At a 30% cache miss rate, 3.6 million requests reach the origin server, each costing $0.008 in compute, for a total origin compute cost of $28,800. Bandwidth adds $15 (12M × 15 KB ÷ 1 GB × $0.085). Bot management tools cost approximately $6,000 per year (Cloudflare Business tier plus custom rule maintenance). Total: $34,815. Under Strategy D, the publisher pre-generates SOM for all articles. At a 2% cache miss rate, only 240,000 requests reach the origin, costing $1,920 in compute. SOM generation costs $5,400 per year. Reduced bot management costs $1,200. Content negotiation edge workers cost $600. Total: $9,124.

Large publisher (50M agent requests/month). Under Strategy B, 600 million annual requests with a 20% cache miss rate produce 120 million origin renders at $0.006 each (reflecting enterprise compute pricing), totaling $720,000 in compute. Bandwidth at volume pricing ($0.06/GB) adds $540. Enterprise bot management costs $120,000. Total: $840,540. Under Strategy D, the 2% cache miss rate reduces origin renders to 12 million at $0.006, costing $72,000. SOM infrastructure costs $156,000 at enterprise scale. Reduced bot management costs $42,000. Edge workers for content negotiation cost $24,000. Total: $294,000.

Figure 2 shows the net annual savings by publisher tier when switching from Strategy B (unstructured) to Strategy D (SOM-first).

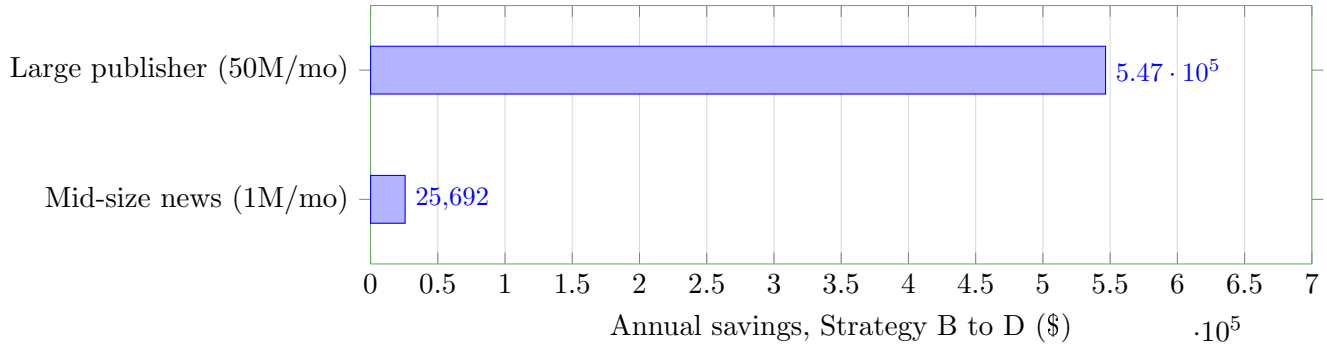

Figure 2: Net annual infrastructure savings when switching from Strategy B (unstructured HTML) to Strategy D (SOM-first with content negotiation). Mid-size tier reflects 74% cost reduction; large tier reflects 65% reduction.

5 Benefit Model

5.1 Direct cost savings

The primary quantifiable benefit of SOM-first serving is infrastructure cost reduction. As detailed in Section 4, mid-size publishers save approximately $25,700 per year and large publishers save approximately $547,000 per year by switching from unstructured to SOM-first serving. These savings are driven primarily by the elimination of origin compute for agent requests, which accounts for 83% of per-request costs under Strategy B.

The savings scale linearly with agent traffic volume, which is itself growing rapidly. At conservative projections (18% annual growth [2]), the mid-size publisher's savings grow to approximately $30,300

in year two and $35,800 in year three, yielding cumulative three-year savings of approximately $91,800.

5.2 Agent comprehension improvement

Beyond cost savings, SOM-first serving improves the quality of agent interaction with publisher content. Our WebTaskBench evaluation [6] measured agent task performance across three web page representations: raw HTML, Markdown, and SOM. Agents operating on SOM representations achieved higher accuracy on content extraction, navigation, and interaction tasks compared to both HTML and Markdown inputs.

The comprehension improvement stems from SOM's preservation of semantic structure. Unlike HTML, which mixes content with visual presentation, or Markdown, which strips structural information, SOM preserves element types (headline, paragraph, image, link), content hierarchy (sections, subsections), interactive affordances (forms, buttons, navigation), and page region semantics (header, main content, sidebar, footer) [20]. This structural preservation enables agents to locate and extract information without the parsing heuristics and error-prone transformations required for HTML processing.

5.3 Attribution and referral potential

Structured representations create the technical foundation for agent-to-publisher attribution. A SOM representation can include canonical URLs, author metadata, licensing terms, and citation preferences in machine-readable fields. When an agent processes SOM content and presents information to a user, it has the structured metadata needed to provide accurate attribution and source links.

While no standardized attribution protocol yet exists, several AI companies have begun licensing publisher content [14, 15] and developing referral mechanisms. Perplexity's Publisher Program [16] provides traffic analytics and citation links to participating publishers. SOM-first serving positions publishers to participate in these and future attribution frameworks by ensuring that agents receive content with clean, parseable metadata.

5.4 Content control and editorial sovereignty

SOM-first serving gives publishers explicit control over what agents receive. Rather than allowing agents to process the full HTML page (including advertising markup, internal analytics, A/B test variants, and navigation elements), the publisher defines the SOM representation to include only the editorial content they intend to share.

This selective disclosure is analogous to the distinction between a website's public pages and its API. The HTML page is the full user experience; the SOM representation is the structured content interface. Publishers can include or exclude specific content elements, add licensing metadata, and version their SOM representations independently of visual redesigns.

5.5 Future-proofing for agent-mediated discovery

As AI agents increasingly mediate how users discover and consume web content, publishers that serve structured representations position themselves for this emerging distribution channel. Agent-mediated discovery differs from search engine optimization in a fundamental way: agents do not merely link to content but process and present it. Publishers whose content is accurately represented in structured form will receive more faithful treatment in agent responses.

The transition from search-mediated to agent-mediated discovery is already underway. Cloudflare reports that AI user-action crawling increased more than 15x in 2025 [1], and major search engines are integrating agent capabilities into their products. Publishers that adopt structured serving early gain experience with the format, build the necessary infrastructure, and establish relationships with agent platforms before the transition reaches critical mass.

Additionally, SOM-first serving provides a direct channel for downstream cost reduction on the agent side. LLM input pricing ranges from $0.15 to $15.00 per million tokens depending on model tier [21, 22]. By reducing input tokens by 4x [5], SOM simultaneously reduces the agent operator's inference cost, creating an incentive for agents to prefer SOM-serving publishers.

6 Sensitivity Analysis

6.1 Traffic growth scenarios

Agent traffic growth directly affects the economic case for SOM adoption, as higher traffic volumes amplify per-request cost differences. We model three growth scenarios based on Cloudflare data [1, 2]:

Conservative (18% annual growth): Extrapolates the overall AI crawler traffic growth rate reported by Cloudflare between May 2024 and May 2025.

Moderate (40% annual growth): Reflects the compound effect of new AI platforms entering the market and existing platforms expanding their crawling scope.

Aggressive (75% annual growth): Weights toward the user-action agent category, which grew more than 15x in 2025 and represents the fastest-growing segment of AI crawling.

Figure 3 illustrates the projected AI bot share of HTML requests under each scenario, with 2023 to 2025 values based on Cloudflare observations and 2026 to 2027 values projected from the three growth rates.

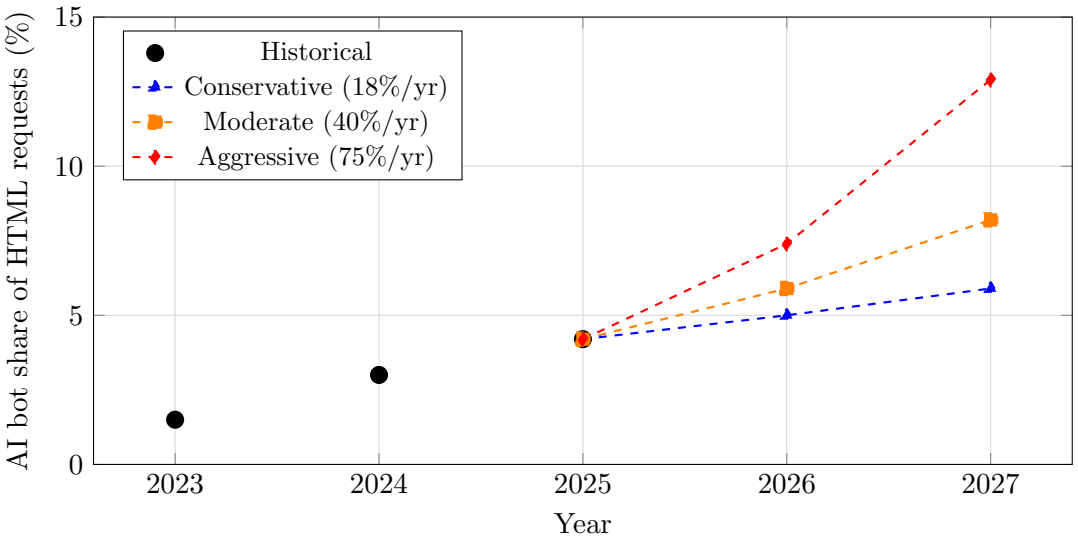

Figure 3: AI bot share of HTML requests, 2023 to 2027. Historical values (solid markers) from Cloudflare Radar data [1, 2]. Projected values (dashed lines) extrapolate three growth scenarios. The 2025 baseline of 4.2% excludes Googlebot's additional 4.5% share, a portion of which also serves AI training.

Under the moderate scenario, AI bot traffic approximately doubles by 2027 relative to the 2025 baseline. Under the aggressive scenario, it triples. Each doubling of agent traffic doubles the absolute cost savings from SOM adoption while leaving percentage savings unchanged.

6.2 Pricing sensitivity

The economic advantage of SOM-first serving depends primarily on the gap between origin compute cost (the dominant expense under Strategy B) and SOM serving cost (largely CDN bandwidth and cached responses under Strategy D). If cloud compute prices decline significantly, the gap narrows.

We model three pricing scenarios and compute the monthly net savings (Strategy B cost minus Strategy D cost) as a function of monthly agent request volume. Figure 4 presents the results.

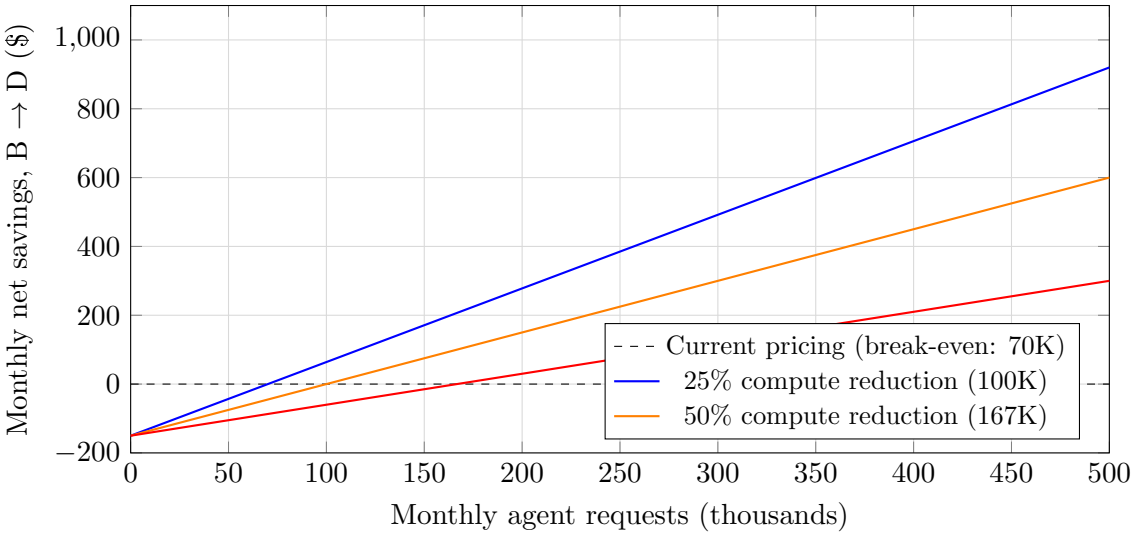

Figure 4: Monthly net savings from switching to SOM-first serving (Strategy D) as a function of agent traffic volume under three compute pricing scenarios. The intersection with the dashed zero line indicates the break-even point. Even with a 50% reduction in compute costs, SOM-first serving yields positive savings above 167,000 monthly requests.

Under current pricing, the break-even point is approximately 70,000 agent requests per month. With a 25% compute price reduction, break-even moves to approximately 100,000 requests. With a 50% reduction, break-even extends to approximately 167,000 requests. Even under aggressive compute price declines, SOM-first serving remains economically justified for publishers receiving more than 200,000 agent requests per month, a threshold that encompasses the majority of commercial publishers.

The break-even calculations use the following linear cost models:

$$C_B = F_B + V_B \cdot R$$
$$C_D = F_D + V_D \cdot R$$

where C_B and C_D are the monthly costs of Strategies B and D respectively, F denotes fixed monthly costs (bot management baseline for B, SOM pipeline for D), V denotes variable cost per thousand requests, and R is the monthly request volume in thousands. The break-even point occurs at $R^* = (F_D - F_B)/(V_B - V_D)$.

6.3 Adoption curves and network effects

SOM adoption exhibits positive network effects. As more publishers serve SOM, agent frameworks have greater incentive to request and process SOM representations. As more agents support SOM, publishers face lower barriers to adoption because the investment reaches a larger audience.

The adoption dynamics resemble those of structured data markup (Schema.org, JSON-LD) in the search engine context. Schema.org adoption was initially slow, driven by a small number of large publishers, but accelerated once major search engines provided tangible ranking benefits for

structured markup. We anticipate a similar trajectory for SOM: early adoption by large publishers with clear economic incentives, followed by broader adoption as agent platforms formalize their preference for structured representations.

The critical threshold is the point at which major agent platforms (ChatGPT, Claude, Perplexity) prefer SOM representations when available. Once this preference is established and communicated through documentation or pricing incentives, publishers face competitive pressure to adopt, as non-adopting publishers risk lower-quality representation in agent responses compared to adopting competitors.

7 Case Studies

We illustrate the cost model with three representative publisher profiles. These are composite profiles based on industry data, not individual organizations.

7.1 News publisher profile

Consider a regional news publisher with 500,000 total pages, publishing 30,000 new articles per month. Total site traffic is 5 million page views per month, of which approximately 1.5 million (30%) are from AI agents.

Under Strategy B, the publisher's annual agent-serving cost is approximately $52,000. Origin compute dominates: at a 30% cache miss rate and $0.008 per render, 5.4 million annual origin renders cost $43,200. Bot management adds $8,000 per year.

Under Strategy D, the publisher pre-generates SOM as part of the editorial workflow at $0.02 per article, totaling $7,200 per year. With SOM served from CDN cache, origin compute drops to $1,700. Reduced bot management costs $2,000. Edge workers cost $1,100. Total annual cost: approximately $12,000, a 77% reduction yielding $40,000 in annual savings.

An additional qualitative benefit: agents receiving SOM correctly extract article headlines, author bylines, publication dates, and section categorization, enabling accurate attribution when citing the publisher's reporting.

7.2 E-commerce catalog

A mid-size e-commerce retailer maintains a catalog of 200,000 product pages with approximately 2,000 price and inventory updates per day. Agent traffic averages 3 million requests per month, driven by comparison shopping agents and product research assistants.

Under Strategy B, dynamic product pages require database queries for real-time pricing and availability, resulting in a higher per-render cost of $0.012. At a 25% cache miss rate, annual compute costs reach $108,000. With bot management at $12,000 per year, total annual cost is approximately $120,000.

Under Strategy D, SOM representations encode structured product data (name, price, availability, specifications, reviews) in a machine-readable format. The publisher implements event-driven SOM regeneration, updating SOM only when underlying product data changes. Annual SOM

infrastructure costs $18,000, origin compute drops to $8,600, and bot management costs $5,000. Total: approximately $32,000, a 73% reduction saving $88,000 per year.

The structured product data in SOM enables shopping agents to perform accurate price comparisons and availability checks without the parsing errors common with HTML scraping.

7.3 Documentation site

A SaaS company maintains a technical documentation site with 5,000 pages, updating approximately 100 pages per week. Agent traffic is 200,000 requests per month, primarily from developer tool agents and coding assistants retrieving API reference documentation.

The site is statically generated, so origin compute costs are negligible. Under Strategy B, the total annual cost of serving agent traffic is approximately $3,000, consisting primarily of bot management overhead.

Under Strategy D, SOM representations are generated at build time, adding 3 minutes to the 15-minute build process at negligible marginal cost. Annual SOM-first costs are approximately $1,500, a 50% reduction driven by simplified bot management. At this scale, the economic case for SOM is modest.

The primary benefit is qualitative: coding assistants receiving SOM correctly identify code examples, API endpoints, parameter types, and return values, producing more accurate responses when developers query the documentation. For a SaaS company, the downstream effect on developer experience and API adoption may outweigh the direct cost savings.

8 Limitations

Several limitations constrain the precision of our estimates. First, our per-request cost figures are derived from published CDN and compute pricing [7, 8, 9], which may not reflect the negotiated rates that large publishers obtain through enterprise contracts. Actual costs for large publishers may be 30 to 50% lower than list pricing, which would reduce absolute savings but not the percentage reduction between strategies.

Second, our cache miss rate estimates (30% for unstructured agent traffic, 2% for SOM) are based on industry reports and analysis of crawler behavior patterns [3]. Individual publishers may experience significantly different cache performance depending on content update frequency, CDN configuration, and the specific agents accessing their sites.

Third, we model SOM generation as a fixed per-page cost of $0.02. In practice, generation complexity varies with page structure, and certain page types (highly interactive applications, single-page applications with client-side routing) may require substantially more effort to represent accurately in SOM.

Fourth, our model does not account for the engineering time required to implement and maintain SOM serving infrastructure. For publishers with limited engineering resources, implementation costs may exceed first-year savings at the mid-size tier.

Finally, the benefit model's attribution and referral estimates are speculative, as no standardized agent attribution protocol yet exists. The economic value of agent-mediated discovery depends on the development of such protocols and their adoption by major AI platforms.

9 Conclusion

The question facing web publishers is no longer whether AI agents will consume their content, but how to serve that content efficiently. Our analysis demonstrates that the current default, serving full HTML pages to agents that need only the semantic content, imposes significant and growing costs on publishers. For a mid-size news site receiving one million agent requests per month, unstructured serving costs approximately $35,000 per year. For large publishers, the figure exceeds $800,000.

SOM-first serving, where publishers generate and serve structured semantic representations through HTTP content negotiation, reduces these costs by 60 to 80% for dynamic sites. The economics are straightforward: SOM representations are smaller (reducing bandwidth), static per content version (enabling near-perfect caching and eliminating origin compute), and standards-based (reducing bot management overhead).

What makes this finding particularly compelling is the alignment between publisher economics and agent performance. SOM is not merely cheaper to serve; it is also the format that agents can most effectively process. Our prior work shows that SOM reduces input tokens by 4x while preserving the semantic content, page structure, and interactive affordances that agents need for accurate comprehension [5, 6]. The format that eliminates 75% of token waste for agents simultaneously eliminates the majority of serving costs for publishers.

The break-even analysis shows that SOM adoption becomes economically justified at approximately 50,000 to 170,000 agent requests per month, depending on compute pricing trends. As agent traffic continues to grow at 18 to 75% annually [2], the threshold for economic justification will encompass an increasingly broad range of publishers.

For publishers evaluating their agent serving strategy, the choice is not between openness and economics. Structured serving offers both.

References

[1] Cloudflare. The 2025 Cloudflare Radar Year in Review. December 2025. https://blog.cloudflare.com/radar-2025-year-in-review/

[2] Cloudflare. From Googlebot to GPTBot: Who's Crawling Your Site in 2025. 2025. https://blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/

[3] Cloudflare. The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals. October 2025. https://blog.cloudflare.com/crawlers-click-ai-bots-training/

[4] HTTP Archive. Web Almanac 2025: Page Weight. January 2026. https://almanac.httparchive.org/en/2025/page-weight

[5] Hurley, D. The Hidden Tax: Quantifying Token Waste in Agent-Web Interaction. Plasmate Labs. March 2026.

[6] Hurley, D. Does Format Matter? Agent Task Performance Across Web Representations. Plasmate Labs. March 2026.

[7] Amazon Web Services. Amazon CloudFront Pricing. https://aws.amazon.com/cloudfront/pricing/

[8] Fastly. Pricing. https://www.fastly.com/pricing

[9] Vercel. Pricing. https://vercel.com/pricing

[10] Cloudflare. Declaring your AIndependence: block AI bots, scrapers and crawlers with a single click. July 2024. https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/

[11] Cloudflare. AI Audit. 2024. https://blog.cloudflare.com/ai-audit/

[12] Fielding, R., Nottingham, M., Reschke, J. HTTP Semantics. RFC 9110. June 2022. https://www.rfc-editor.org/rfc/rfc9110

[13] Koster, M., Illyes, G., Zeller, H., Sassman, L. Robots Exclusion Protocol. RFC 9309. September 2022. https://www.rfc-editor.org/rfc/rfc9309

[14] OpenAI. OpenAI and Associated Press expand partnership. July 2023.

[15] OpenAI. Axel Springer and OpenAI partner to deepen beneficial use of AI in journalism. December 2023.

[16] Perplexity. Perplexity Publisher Program. 2024. https://www.perplexity.ai/hub/blog/perplexity-s-publisher-program

[17] The New York Times Company v. Microsoft Corporation et al. Case No. 1:23-cv-11195 (S.D.N.Y.). December 2023.

[18] Cloudflare. Workers Pricing. https://developers.cloudflare.com/workers/platform/pricing/

[19] Akamai. State of the Internet Report: Web Traffic and Bot Activity. 2025. https://www.akamai.com/internet-station/cyber-attacks/state-of-the-internet-report

[20] Hurley, D. Semantic Object Model: A Structured Representation for Agent-Web Interaction. Plasmate Labs. March 2026.

[21] OpenAI. API Pricing. March 2026. https://openai.com/api/pricing/

[22] Anthropic. API Pricing. March 2026. https://www.anthropic.com/pricing