# The Hidden Tax: Quantifying Token Waste in Agent-Web Interaction

David Hurley
Plasmate Labs

March 2026

**Abstract**

AI agents that browse the web consume raw HTML as input to large language models. We estimate that approximately 75% of the tokens in a typical web page encode visual presentation, tracking, and layout information that is irrelevant to agent reasoning. Using public data from Cloudflare Radar (crawl volumes), HTTP Archive (page sizes), provider pricing pages, and our own WebTaskBench measurements (token efficiency), we estimate the annual economic cost of this waste. A bottom-up model based on agent user counts yields \$1B to \$3B per year. A top-down model calibrated against Cloudflare traffic data yields \$1B to \$5B per year. We survey 10 major agent frameworks and find that none use structured semantic representations by default: the three largest orchestration frameworks (LangChain, LlamaIndex, CrewAI) default to plain text extraction, while dedicated scraping tools default to Markdown. Both approaches reduce waste partially but neither eliminates it. Structured representations such as SOM (Semantic Object Model) would eliminate approximately 75% of token waste while preserving semantic content, page structure, and interactive affordances.

## 1 Introduction

Every AI agent that browses the web pays a hidden tax. When an agent fetches a web page and passes it to a language model, the majority of the input tokens encode information that is irrelevant to the agent's task: CSS class names, inline styles, tracking pixels, advertising markup, and layout containers.

This waste has three costs. First, it consumes API budget. At \$0.50 to \$3.00 per million input tokens, noise tokens translate directly into money spent on nothing. Second, it displaces useful content from the context window, limiting how many pages an agent can process in a single pass. Third, it increases inference latency, as the model must process tokens that contribute nothing to the answer.

Despite the practical significance of this waste, we are not aware of any prior work that quantifies it at an industry level. This paper presents two complementary estimation approaches and a survey of current mitigation practices across major agent frameworks.

## 2 Data Sources

All inputs to our model are from public, verifiable sources.

### 2.1 Agent crawl volume

Cloudflare Radar's 2025 Year in Review [1] reports that Cloudflare handles approximately 81 million HTTP requests per second on average. AI bots account for 4.2% of HTML request traffic, with an additional 4.5% from Googlebot (which serves both search indexing and AI training). AI "user action" crawling (agents performing tasks directed by users) increased by over 15x in 2025, making it the fastest-growing category of AI crawling.

Cloudflare's "crawl-to-click gap" analysis [2] provides further detail: training drives approximately 80% of AI crawling, while user-action and search crawling account for the remainder.

## 2.2 Web page sizes

The HTTP Archive Web Almanac 2025 [3] reports a median mobile homepage size of 2.56 MB (total transfer) and 2.86 MB on desktop. For agent consumption, the relevant metric is rendered HTML document size, which we measured at an average of 278 KB across the 50 websites in WebTaskBench [4].

## 2.3 Token efficiency

Our WebTaskBench evaluation [4] measured token consumption across three web page representations on 50 real websites:

| Format | Mean input tokens | Ratio to HTML |
|---|---:|---:|
| Raw HTML | 33,181 | 1.0x |
| SOM (Semantic Object Model) | 8,301 | 4.0x fewer |
| Markdown (text extraction) | 4,542 | 7.3x fewer |

Table 1: Average input tokens per page by representation format.

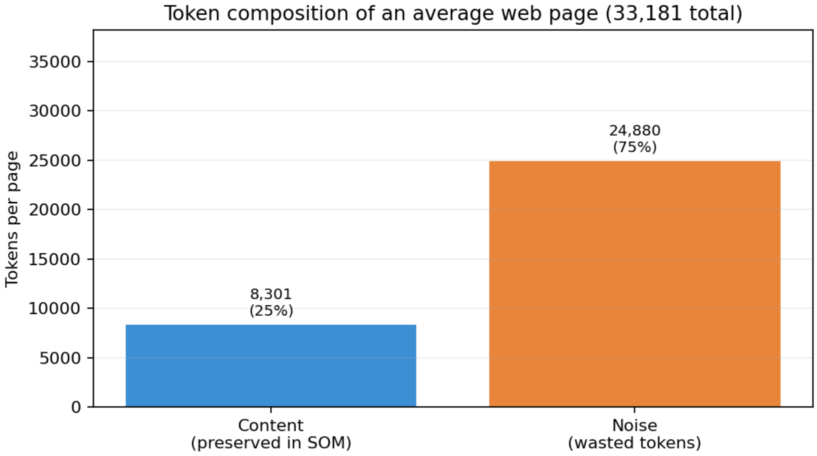

Figure 1: Token composition of an average web page.

The difference between HTML and SOM (24,880 tokens per page) represents tokens that encode visual presentation but not semantic content. SOM preserves page structure, element types, interactive affordances, and content hierarchy while eliminating this noise. We treat the SOM token count as the "useful content" baseline and the difference as waste.

## 2.4 LLM API pricing

We use representative input prices from provider pricing pages as of March 2026:

| Model | Input price ($/M tokens) |
|---|---:|
| GPT-4o | 2.50 |
| GPT-4o Mini | 0.15 |
| Claude Sonnet 4 | 3.00 |
| Claude Opus 4.6 | 15.00 |
| Gemini 2.5 Pro | 1.25 |

Table 2: Representative LLM input pricing (March 2026).

We estimate a weighted average of $0.75 per million input tokens for agent web browsing workloads, assuming a mix skewed toward mid-tier and budget models.

# 3 Estimation: Bottom-Up

Our primary estimate builds upward from agent user counts.

## 3.1 Agent browsing volume

We estimate daily LLM-consumed web page fetches from public reports on agent platform usage:

| Platform | Est. users | Browse rate | Pages/day |
|---|---:|---:|---:|
| ChatGPT (web browsing) | 60M | 5 pg/session, 3x/wk | 129M |
| Claude (web search) | 7.5M | 5 pg/session, 2x/wk | 11M |
| Perplexity | 20M | 5 pg/query, 5x/wk | 71M |
| Other agents | 10M instances | 20 pg/day | 200M |
| **Total** | | | **411M** |

Table 3: Estimated daily LLM-consumed page fetches by platform.

## 3.2 Token waste calculation

Per page, 24,880 tokens are waste (HTML minus SOM). After adjustments for current preprocessing:

- 45% of agents use Markdown preprocessing, reducing waste by approximately 70% for those agents
- 30% of agents truncate HTML, reducing waste by approximately 30%
- 15% of fetches are cache hits (no LLM consumption)

Effective waste per page after adjustments: 13,323 tokens.

Annual waste: 411M pages/day × 13,323 tokens × 365 = $2.0 \times 10^{15}$ tokens/year.

Annual cost at $0.75/M tokens: **$1.5 billion/year**.

Sensitivity: at $0.50/M, the estimate is $1.0B. At $1.50/M, it is $3.0B.

# 4 Estimation: Top-Down

As a cross-check, we derive an estimate from Cloudflare traffic data.

Cloudflare reports 81M requests/second and AI bots at 4.2% of HTML traffic. Assuming Cloudflare handles approximately 25% of global web traffic, and that user-action agents represent 0.5% of total AI crawling (consistent with user-action being the smallest but fastest-growing category), we estimate approximately 1.5 billion user-action page fetches per day globally. With 25% LLM consumption and the same waste adjustment, this yields approximately $5B/year at $0.75/M tokens.

The top-down estimate is higher because it captures enterprise and custom agents not included in the bottom-up platform counts. The two approaches bracket the estimate: **$1B to $5B per year**.

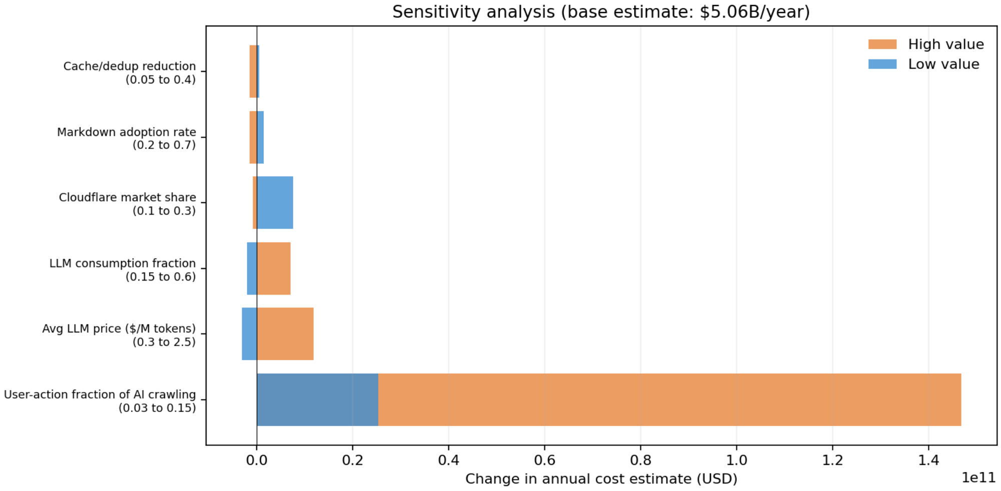

Figure 2: Sensitivity analysis: impact of parameter variation on annual cost estimate.

## 5 Framework Survey

We surveyed the default web page loading behavior of 10 major agent frameworks and tools to assess current waste mitigation practices.

| Framework | Default format | Method |
|---|---|---|
| LangChain | Plain text | BeautifulSoup get_text() |
| LlamaIndex | Plain text | BeautifulSoup get_text() |
| CrewAI | Plain text | BeautifulSoup get_text() |
| Crawl4AI | Markdown | Custom HTML-to-Markdown |
| Firecrawl | Markdown | Readability + Markdown |
| Jina Reader | Markdown | Custom extraction |
| AutoGPT | Markdown | Delegates to Jina/Firecrawl |
| Browser Use | Custom DOM | Accessibility tree + DOM |
| Stagehand | Custom DOM | Accessibility tree |

Table 4: Default web page representation by framework.

No framework uses a structured semantic representation by default. The three largest orchestration frameworks (LangChain, LlamaIndex, CrewAI) default to plain text extraction via `BeautifulSoup.get_text()`, which strips HTML tags but also discards all semantic structure. Dedicated scraping tools default to Markdown, which preserves some text hierarchy but loses element types, interactive affordances, and page region structure.

Browser automation frameworks (Browser Use, Stagehand) use custom DOM and accessibility tree representations for action selection, but these are not standardized, not shared across frameworks, and not designed for LLM content comprehension.

# 6 Discussion

## 6.1 The waste is real and quantifiable

Our estimates, derived from independent data sources using complementary methodologies, converge on a range of $1B to $5B per year in token waste from HTML presentation noise. This represents a meaningful fraction of the total LLM API market.

## 6.2 Current mitigation is partial

Markdown extraction reduces waste by approximately 70% but eliminates semantic structure. Plain text extraction (the default in major frameworks) reduces waste similarly but produces even less structured output. Neither approach preserves element types, interactive affordances, or page regions.

## 6.3 Structured representations would eliminate the waste

SOM reduces input tokens by 4.0x versus raw HTML while preserving all semantic content, page structure, element roles, and interaction affordances. If adopted broadly, SOM would reduce annual agent token waste by approximately 75%, saving $0.75B to $3.75B per year at current pricing.

As LLM prices decrease, the absolute dollar savings will decline, but the efficiency gains (fewer tokens, faster inference, more content per context window) persist regardless of pricing.

# 7 Limitations

Agent user counts and browsing rates are estimated from public reports and may not reflect actual usage. The Cloudflare data covers traffic through their network only. Our token efficiency measurements are from 50 websites and may not be representative of all web content. LLM pricing changes frequently.

The framework survey captures default behavior only. Individual deployments may override defaults with custom preprocessing.

# 8 Conclusion

AI agents collectively spend $1B to $5B per year processing HTML presentation noise that is irrelevant to their tasks. This waste is not an inherent cost of web browsing but an artifact of using a presentation format (HTML) for a comprehension task (LLM reasoning). Structured semantic representations like SOM would eliminate approximately 75% of this waste while providing richer, more reliable input for agent reasoning.

# References

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

[2] 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/

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

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