Measuring the Citation Gap

Retrieval Eligibility and Extraction Readiness on Small and Mid-Sized Business Websites Across Four Countries

Vikas

vSourceCode

August 2026

Measuring the Citation Gap

Retrieval Eligibility and Extraction Readiness on Small and Mid-Sized Business Websites Across Four Countries

Vikas · vSourceCode · 18 August 2026

Abstract

Retrieval-grounded generative systems can retrieve web sources and use them to construct answers, including answers that expose source citations. Existing measurement of web quality addresses the rendered document — accessibility audits execute JavaScript, structured-data surveys sample crawler corpora, and consent studies examine robots.txt in isolation. None measures what a retrieval client actually receives when it requests an ordinary business homepage and is not a browser. This paper specifies a protocol for that measurement and applies it to 428 small and mid-sized business websites across four countries and four sectors. Each domain is requested under up to four conditions with every attempt recorded, so refusals are reported rather than converted into missing data, and seven binary preconditions are observed with the supplying observer recorded per value. We then define the **Citation Gap** operationally as the set of sites that satisfy the access preconditions — reachable, permitted, indexable — but do not satisfy all four extraction-readiness preconditions specified by the protocol. Of 349 access-eligible sites, 187 (53.6%) fall below that threshold. Its structure is unexpectedly shallow: 114 of the 187 fail exactly one extraction precondition, most often a top-level heading (45 sites), structured data (38), or a meta description (31), and none fails solely on non-JavaScript content. Adoption of llms.txt, the convention introduced for this purpose, is 20% overall and is associated with sites already outside the gap (29% versus 14%). Blocking of AI crawlers in robots.txt is rare (13 of 382 measurable sites); the observed access failures occur instead at the network edge. An independent local implementation of Google’s agentic accessibility-tree audit agrees with it on 55% of sites, so that measure is reported as implementation-specific. The sample is designed to exercise the protocol across heterogeneous business websites, not to estimate population prevalence. No engine behaviour is observed and no causal claim is made.

JEL classification: L86, M31, O33, L26

Suggested citation. Vikas (2026). *Measuring the Citation Gap: Retrieval Eligibility and Extraction Readiness on Small and Mid-Sized Business Websites Across Four Countries.* vSourceCode Working Paper, August 2026.

Canonical version: https://vsourcecode.com/reports/book/measuring-the-citation-gap-research-paper

Conflict of interest. The author operates vSourceCode, which provides technical consulting to businesses on the subjects this paper measures. This paper recommends no product or service, links to no commercial offering of the author’s, and specifies a self-assessment protocol (Appendix A) requiring only free tools operated by third parties. The reader should nonetheless weigh the finding that many audited sites fall short of a readability standard against the author’s commercial interest in such shortfalls being recognised.

1 · INTRODUCTION

A second retrieval layer has appeared between the reader and the page. Retrieval-grounded generative systems can fetch web sources and use them to construct an answer, and several expose citations to the sources used. Adobe reported AI-referred traffic in the United States growing roughly twelvefold between July 2024 and February 2025.[12] Google states that AI Overviews reach two billion users a month.[14] Alphabet reported the Gemini application at 950 million monthly actives in its second-quarter 2026 results.[15] Anthropic's documentation states that citations are always enabled for its web search tool, returning a source URL, title, and snippet for every result.[16]

Whether the pages being retrieved are constructed so that this can happen is a separate question, and it has not been measured in the form that matters. The obstacle is not neglect but instrumentation: the major surveys of web quality all observe the *rendered* document, and a retrieval client is not a browser.

This paper specifies a protocol for observing what a non-rendering, honestly identified client receives, and applies it to 428 small and mid-sized business websites in the United States, the United Kingdom, Canada, and the Netherlands, across home services, legal services, medical and dental practices, and a mixed small-business group.

Its central contribution is an operational definition. *Citation Gap* is used loosely in practitioner writing to mean that AI visibility differs from search visibility. That is a slogan, not a measure. Section 7 defines it as a specific, computable partition of a measured sample — the sites an engine can reach and is permitted to use, but which fall below the protocol's extraction-readiness threshold — and reports its size, depth, and composition.

Three results follow. First, the gap holds 53.6% of access-eligible sites, so it is where the shortfall concentrates. Second, it is shallow: 114 of 187 gapped sites fail exactly one of four extraction preconditions. Third, the newest convention is not the binding constraint — llms.txt adoption is twice as high among sites already outside the gap as among sites inside it.

A separate result concerns where access failures occur. Blocking of AI crawlers by name in robots.txt is rare, but 70 sites failed a first request that a permitted crawler would have made, predominantly through invalid TLS certificates, request filtering, and server errors. Four sites publish a robots.txt permitting every crawler and then refuse one at the edge.

The claim is deliberately limited. Satisfying these preconditions is associated with the protocol's operationalisation of citation readiness. It does not establish that any engine retrieves, selects, or names any site, and nothing measured here observes engine behaviour.

2 · RELATED WORK, AND WHAT IT DOES NOT MEASURE

Five literatures bear on this question. Each measures something adjacent, and the gap between them is where this protocol sits.

Generative engine optimisation. Aggarwal et al. introduced GEO at KDD 2024, formalising generative engines as systems that retrieve documents and synthesise a cited answer, and testing nine content modifications against a benchmark of queries.[1] Their contribution is on the *content* margin: given that a source is in the retrieval pool, which textual properties raise its visibility in the generated answer. Subsequent work extends this — Kumar and Lakkaraju on adversarial manipulation of product visibility,[3] Wan et al. on what evidence language models find convincing,[4] Chen et al. on optimisation at scale.[2] All of it assumes the document has been retrieved. None measures whether an ordinary business website enters the pool at all, and the GEO literature is explicit that it operates on sources already available to the engine.

Web accessibility measurement. The WebAIM Million is the largest recurring automated survey of web quality, evaluating the top one million home pages annually; the February 2026 edition detected WCAG 2 failures on 95.9% of pages, averaging 56.1 errors, reversing six years of gradual improvement.[5] It is directly relevant, because the accessibility tree is the structure an automated client traverses. But WAVE analyses **the rendered DOM after scripting and styles are applied**.[5] That is the correct instrument for its research question — a human using assistive technology runs a browser — and the wrong one for ours. A page that renders entirely client-side scores identically to a server-rendered page under WebAIM's method and returns nothing to a retrieval client that does not execute JavaScript. The distinction is not hypothetical: 12% of sites in this sample return no readable content without scripting.

Crawler consent and robots.txt. Sun, Zhuang and Giles conducted the first large-scale study of robots.txt.[7] Longpre et al. produced the definitive modern treatment, auditing consent protocols across 14,000 domains and documenting a rapid rise in AI-specific restrictions, with roughly 28% of the most actively maintained sources in C4 becoming fully restricted within a

single year, alongside systematic inconsistencies between what sites state in their terms of service and what they express in robots.txt.[6] Two features of that work shape ours. It samples the domains underlying training corpora — high-traffic, editorially managed, disproportionately publishers — and it measures the *stated* policy. This paper samples ordinary small businesses that appear in no training-corpus ranking, and measures stated policy against deployed behaviour by issuing the request. The 4 sites here that permit in writing and refuse in practice are the small-business analogue of the inconsistency Longpre et al. found between terms of service and robots.txt, observed one layer lower in the stack.

Structured data on the web. Large-scale surveys of schema.org adoption, principally through Web Data Commons extraction from Common Crawl, establish prevalence at web scale. Baack's analysis of Common Crawl as a training-data source documents what such corpora systematically include and omit.[8] Crawler-corpus sampling inherits the crawler's own access biases: a site that refused the crawler is absent from the corpus and therefore absent from the survey, which makes these methods structurally unable to measure refusal. A refusal is exactly what this protocol is built to record.

Retrieval-augmented generation. The mechanism by which a fetched document becomes a cited claim is well established in the RAG literature following Lewis et al.[9] That work is a systems literature, concerned with retrieval and grounding inside the model pipeline. It takes a corpus as given. The question of which real-world documents are *constructible into* such a corpus is not addressed there, and is the question here.

2.1 A note on crawler-identity attribution

A sixth strand — traffic studies attributing request volume to named AI crawlers — is excluded from the comparison above, for a reason worth stating, since it also explains a design choice in section 4.

The following are first-party observations from a single domain, offered to illustrate a measurement hazard rather than as evidence about the web.

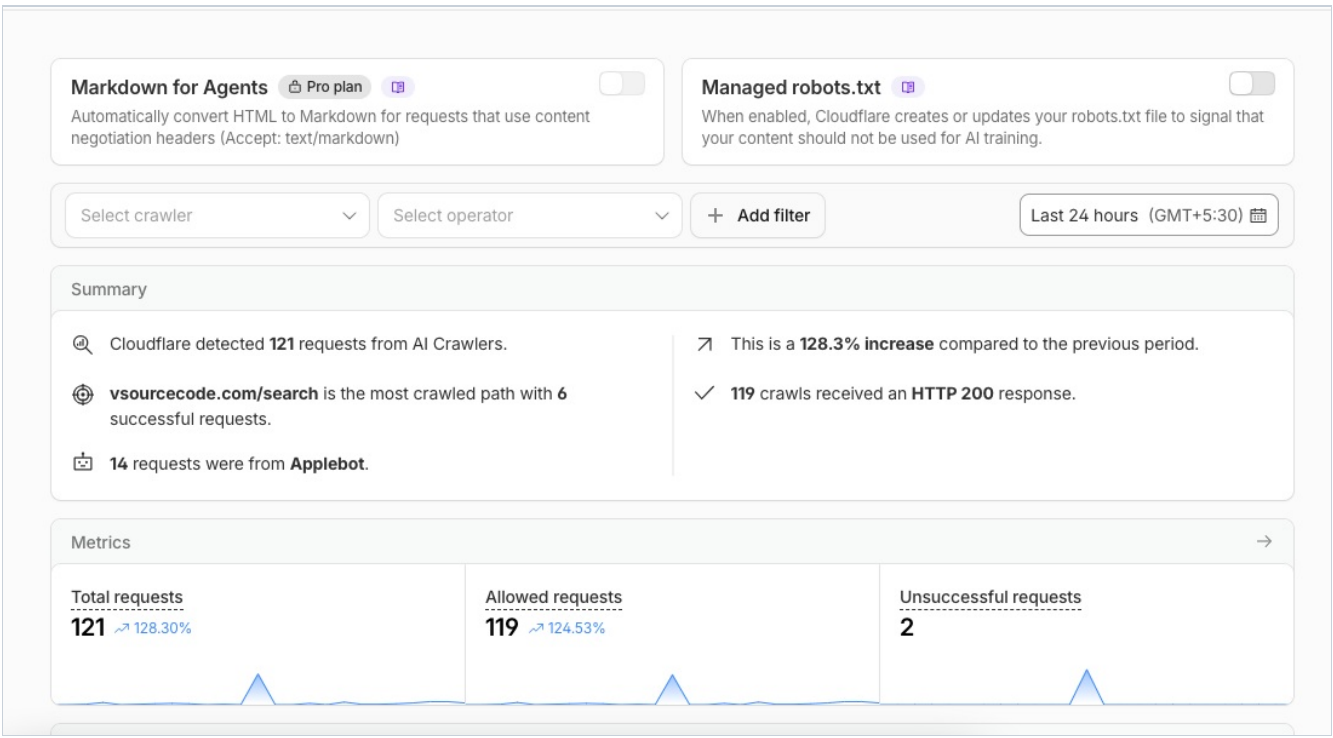

Figure 1. AI-crawler request summary: 121 requests in 24 hours, 119 returning HTTP 200. Collected from the Cloudflare AI Crawl Control dashboard for vsourcecode.com, 24-hour window ending 18 August 2026, GMT+5:30. Single-domain observation; not generalised.

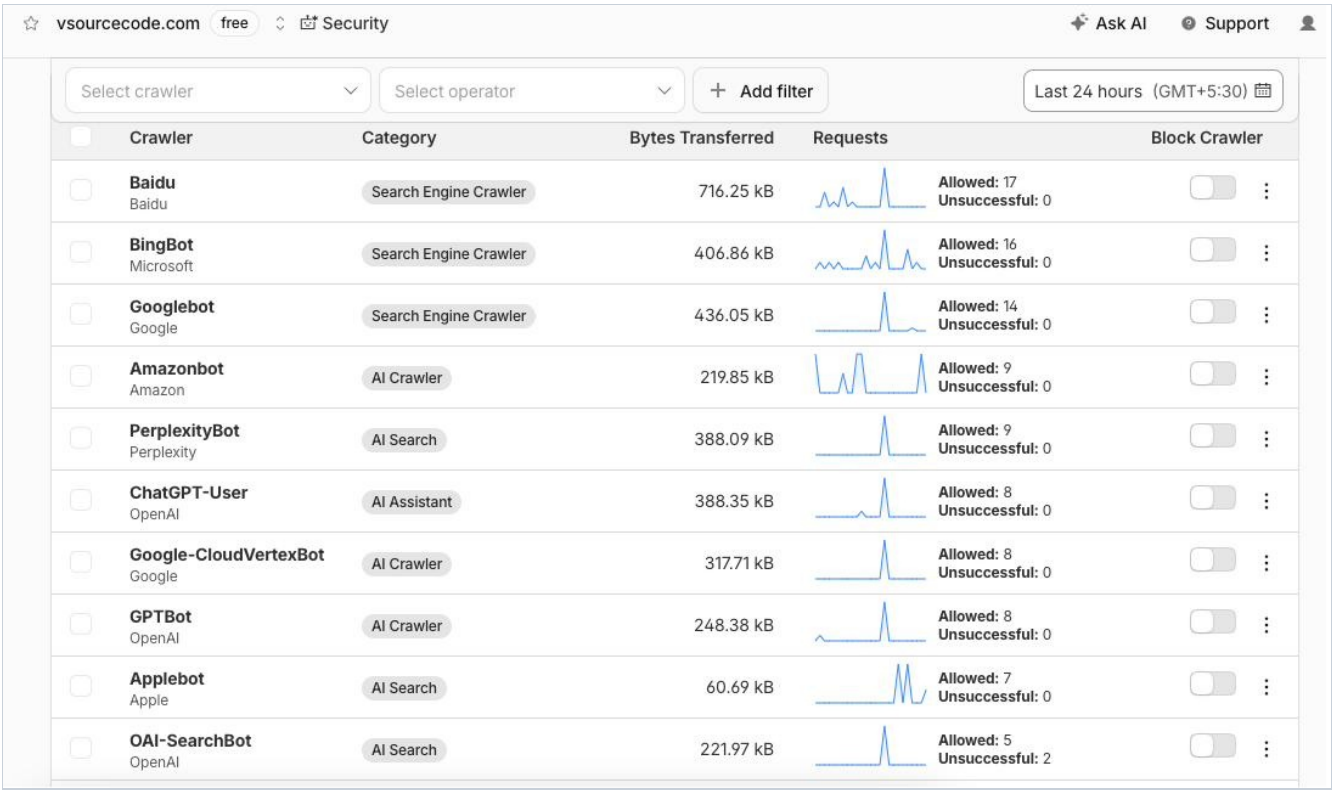

Figure 2. Per-crawler request counts and byte volumes, separated by the dashboard's own categories. Same source and window as Figure 1. Single-domain observation; not generalised.

The category distinction in Figure 2 matters for any such study. Baidu, Bing and Googlebot are indexing crawlers; Amazonbot, GPTBot and Google-CloudVertexBot are collection crawlers; PerplexityBot, Applebot and OAI-SearchBot serve retrieval at query time; ChatGPT-User is an assistant fetching a page because a person asked something. These are different mechanisms with different implications, and a robots.txt written without the distinction governs them indiscriminately.

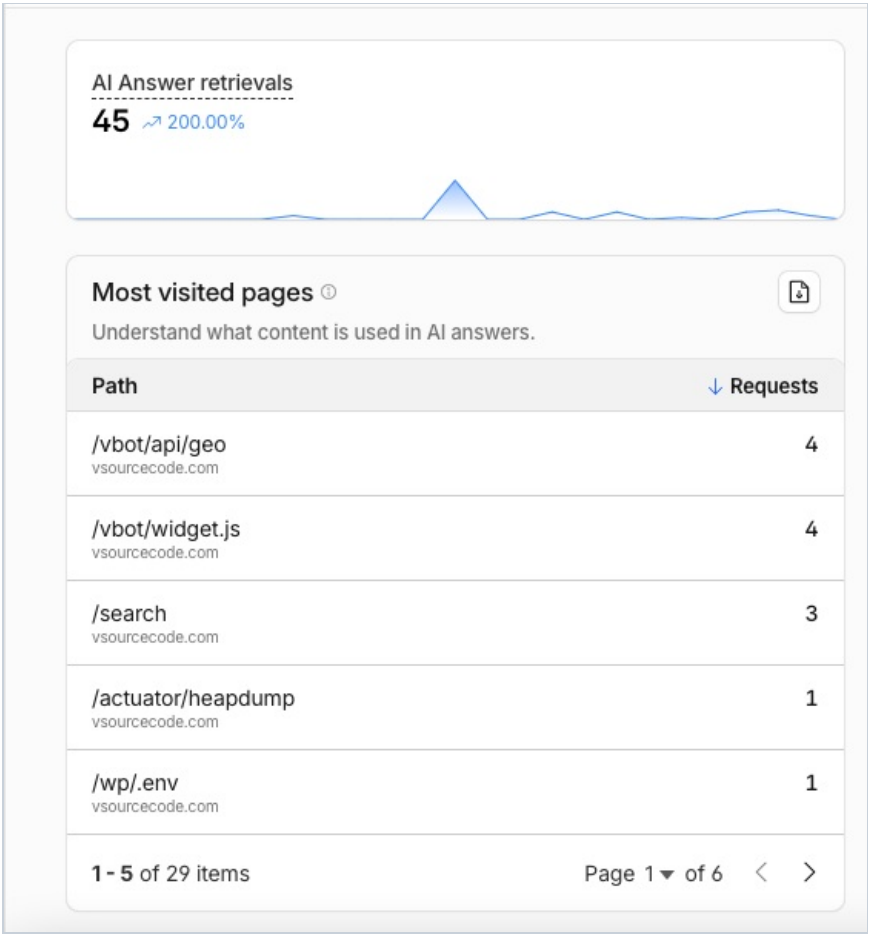

Figure 3. Answer-time retrievals and most-requested paths. Same source and window as Figure 1. Single-domain observation; not generalised.

Figure 3 carries the hazard. Among the most-requested paths attributed to AI clients are `/actuator/heapdump` and `/wp/.env` — a Spring Boot memory-dump endpoint and a WordPress environment file. Neither is a path any answer-retrieval client would

request. They are credential probes arriving under user-agent strings claiming to be AI crawlers.

The user-agent header is self-declared and unverified. Any study attributing volume to named AI crawlers on that basis alone counts impersonators alongside legitimate clients, and the impersonators are not neutral noise: they are hostile requests whose inclusion inflates apparent AI crawl volume. Cryptographic verification of crawler identity exists in draft and is not yet widely deployed. Until it is, user-agent-attributed figures in this area — including those above — should be read as upper bounds.

This is why the protocol specified below makes no use of crawler-identity claims. It measures what a page returns to a client that identifies itself honestly, which is a fact about the page rather than a claim about the requester.

The gap. Existing work measures the rendered page, the stated policy, the crawled corpus, or the content already retrieved. None issues a request as a non-rendering identified client to an ordinary business website and records what comes back. That measurement is the contribution, and the Citation Gap defined in section 7 is what it makes computable.

Two smaller contributions follow from the instrument rather than the sample. The four-condition retrieval ladder (section 4.1) separates refusal from unmeasurability, which single-request instruments cannot do. And the divergence in section 8 between two implementations of the same accessibility-tree construct is, to the author’s knowledge, the first reported comparison of its kind, and bears on any study using that audit as a dependent variable.

---

3 · THE CONSTRUCT

For a retrieval-grounded system to cite a source, at minimum three operations must succeed: it must be permitted to request the page, it must receive something back, and it must be able to locate within that response the claim it intends to attribute. Failure at any of the three ends the sequence.

Citation readiness is the set of technical preconditions under which those operations are possible. Seven properties are measured, each independently observable, each binary, each verifiable by a third party from the URL alone:

	Precondition	Layer
1	No robots.txt rule disallowing any named AI crawler, and no blanket disallow	Access
2	No `noindex` directive in the robots meta tag	Access
3	Readable text present before any script executes	Extraction
4	An `h1` element on the page	Extraction
5	JSON-LD or schema.org microdata present	Extraction
6	Both a title and a meta description	Extraction
7	A conforming llms.txt at the domain root	Declaration

The layer assignment is not decorative; section 7 builds the Citation Gap on it. Access preconditions govern whether a request succeeds. Extraction preconditions are the structural elements this protocol treats as constituting extraction readiness; they are an operational specification, not a claim about what any engine requires. The declaration layer is a single emerging convention, held separate because no engine is documented as requiring it.

The composite count is unweighted. No weighting can be justified from this data: nothing here establishes the relative contribution of any property to any outcome, and a weighted index would embed a claim the measurement cannot support. The count is a descriptive summary, not a model.

Three things the construct is not.

Not a measure of citation. Nothing here observes any engine retrieving, selecting, or naming any site. Sites scoring low exhibit fewer measured preconditions associated with this protocol’s operationalisation of citation readiness — which is a statement about the sites, not a prediction about engines.

Not a ranking factor. Google’s agentic-browsing category, used in section 8 as an external check, is documented as experimental, under development, and not a confirmed ranking factor.[17]

Not a quality measure. A site can satisfy all seven properties and be worthless, or fail several and be the best source on its subject.

---

4 · INSTRUMENT AND METHOD

4.1 Retrieval

Each domain was resolved over DNS, then requested over HTTPS with a user agent identifying the research client and carrying a contact URL. Three resources were retrieved: the homepage, `/robots.txt`, and `/llms.txt`. No JavaScript was executed. Rate-limit responses were honoured with a back-off, and `Retry-After` respected where supplied.

Where the identified request failed, up to three further conditions were attempted and every attempt recorded:

```
c1_agent     identified research UA, https, host as listed   ← headline
c2_browser   common browser UA, SAME url — UA is the only variable
c3_hostswap  browser UA, www./apex flipped
c4_http      browser UA, plain http — TLS-level failures only
```

Only the first two differ by client identity alone. The third and fourth change the origin and are reported separately: a site reached only over plain HTTP has not demonstrated that it serves crawlers, it has demonstrated that its TLS configuration is broken.

This ladder is what distinguishes a recorded refusal from lost data. A single-request instrument records a 403 as absence of data; here, 26 sites that failed the first request were reached by a later condition and remain in every table.

4.2 Parsing robots.txt

RFC 9309 specifies matching on the product token, with the most specific matching group taking precedence.[10] Version suffixes are stripped from both sides before comparison. A permissive prefix reading was computed in parallel as a separate count:

```
// RFC 9309 matches on the product TOKEN, and crawlers take the most specific
// group that matches. Exact-string lookup missed "User-agent: Claude" (which
// ClaudeBot obeys) and "User-agent: GPTBot/1.0". Every one of those misses
// pushed the same way — understating how many sites block AI — so the headline
// was biased, not merely noisy. Longest matching token wins.
function findGroup(parsed, agent, loose) {
  const norm = k => k.split('/')[0].trim();
  const A = norm(String(agent).toLowerCase());
  for (const k of Object.keys(parsed.groups)) {
    if (k !== '*' && norm(k) === A) return { key: k, group: parsed.groups[k] };
  }
  if (!loose) return null;
  /* PREFIX MATCHING IS A SEPARATE, ARGUABLE QUESTION, so it gets its own count
     rather than being folded into the headline. */
```

`Applebot-Extended` governs training-data collection; `Applebot` governs search indexing. They are different agents, and a rule written for one does not govern the other. An instrument that substring-matches reports sites as blocking crawlers they permit. Both readings returned 13 sites here, so no finding depends on the resolution — but that identity is a result, not an assumption.

4.3 File validity, and the difference between absent and unmeasured

Many sites return their homepage for any missing path, so a naive presence check reports a `robots.txt` or `llms.txt` on sites that have neither. A file was counted present only where the response returned HTTP 200, the body was not HTML, and the content was semantically valid for its type:

```
// "absent" and "not measured" were the same empty cell in the 16 Aug run — 78
// rows of unreadable llms.txt sat next to a published 21% adoption figure.
if (!r.ok) return { status: r.status === 404 ? 'absent' : 'unreadable',
                    valid: 0, measured: r.status === 404 ? 1 : 0 };
if (looksHtml(r)) return { status: 'soft404_html', valid: 0, measured: 1 };
const h1    = /^#\s+\S/m.test(b) ? 1 : 0;
const links = /https?:\/\/\S+/.test(b) ? 1 : 0;
const long  = b.length > 100 ? 1 : 0;
const valid = (h1 && links && long) ? 1 : 0;
```

A 404, and an HTML body served at these paths, both establish absence — an observation. A 403 or timeout establishes nothing and is recorded as unmeasured.

4.4 Unmeasurable is not zero

The most consequential design decision is how to treat a property that could not be observed. A site that refuses the request has not been shown to lack a heading; nobody was permitted to look. Scoring it zero manufactures a finding out of a refusal.

Each property carries its value *and the observer that supplied it*, and properties with no observer are excluded from that site's denominator:

```
const pick = (mine, psiVal) => mine !== null && mine !== undefined
  ? { v: mine, s: 'agent' }
  : (psiUsable && psiVal !== null && psiVal !== undefined) ? { v: psiVal, s: 'psi' }
  : { v: null, s: 'na' };

const B = {
  crawlable:   pick(parsed ? ((!star && blocked.length === 0) ? 1 : 0)
                           : (robotsAbsent ? 1 : null), null),
  indexable:   pick(own ? (H.noindex ? 0 : 1) : null, pv('psi_indexable')),
  server_html: pick(own ? ((H.words >= 100) ? 1 : 0) : null, null),
  has_heading: pick(own ? ((H.h1 >= 1) ? 1 : 0) : null, null),
  has_schema:  pick(own ? ((H.jsonld > 0 || H.microdata_schemaorg) ? 1 : 0) : null, null),
  has_meta:    pick(own ? ((H.title && H.meta_desc) ? 1 : 0) : null, /* … */),
  llms_txt:    pick(L.measured ? (L.valid ? 1 : 0) : null, null)
};
```

Denominators therefore differ by property and by site, and every table states its own.

4.5 Second observer

Google PageSpeed Insights was queried for every site. Where the instrument was refused but Lighthouse was not, indexability and title/meta-description were taken from Lighthouse and recorded as such. Sixteen sites carry a Lighthouse-sourced value, which is why those two rows have denominator 400 and the rest 384.

Where both observers answered, the instrument's value is used: the non-rendering view is the study's subject, and the rendered view would flatter it.

Three properties have no Lighthouse equivalent and remain unmeasurable when the page is refused. Lighthouse executes JavaScript and cannot report the non-rendering view; no audit asserts that an `h1` exists; and the structured-data audit is manual, returning no machine-readable result. Lighthouse never exposes robots.txt contents, so crawler blocking cannot be measured that way at all. Where Lighthouse followed a redirect to a different registrable domain (29 sites), its values were withheld.

4.6 Parked-domain detection

Domain-parking pages serve their own `llms.txt` and structured data, and they resolve, respond, and parse cleanly. Left in the sample they inflate every adoption rate. Parking is detected at two layers, body pattern and final-URL host, after a case in which a parked page carrying a law firm's former name entered an earlier run and scored three of seven.

4.7 Deliberate exclusions

Performance metrics were collected and are not reported: repeat measurement minutes apart moved timing figures substantially while structural checks returned identical values. The agentic category's pass-ratio score is not reported — it rewards sparse pages, and a near-empty document scores well. Its WebMCP audits measure enrolment in a Chrome origin trial, effectively zero across a sample of this kind.

---

5 · STUDY DESIGN AND SAMPLE

The sample is designed to exercise the protocol across heterogeneous business websites, not to estimate population prevalence. This is a deliberate choice of research question. A prevalence estimate would require a probability sample of a defined population of business websites, and no sampling frame for that population exists — there is no register of the world's small businesses with websites. What can be done, and what is done here, is to run the protocol across sites that vary systematically in country, sector, firm size, and technology stack, and to establish that the measure discriminates among them, that its failure modes are recoverable, and that its denominators behave as designed.

Every proportion reported below is therefore a property of this sample. None is an estimate of any wider population, and no confidence interval is reported, because the sampling design does not support one.

Domains were drawn from public directories, professional registers, and trade listings, with sector and country recorded at collection. The frame was compiled at approximately 30 domains per cell.

Step	Domains
Unique domains on the source list	484
Less: domain no longer resolves	−37
Domains audited	447
Less: parked domain	−16
Less: instrument failure	−3
Sites analysed	**428**

Two conditions removed a business on grounds of the source list rather than any property of its website: the domain did not resolve, or it served a parking page. The 37 non-resolving domains are 7.6% of a frame drawn from live public directories. Per-domain status is in Appendix C.

Three sites were lost to failure of the instrument rather than behaviour of the site. They are reported rather than dropped quietly, and are the only exclusions here that are not properties of the sample.

No site was excluded for being slow, small, foreign, or unreachable. Where a domain resolved and refused the request, it was retained, because a refusal is itself an observation.

	Home services	Legal	Medical & dental	Mixed small business	Total
US East	29	23	27	23	**102**
US West	27	24	27	0	**78**
United Kingdom	26	19	25	8	**78**
Canada	26	30	25	17	**98**
Netherlands	22	26	24	0	**72**
Total	**130**	**122**	**128**	**48**	**428**

The United States is split East and West as separate collection cells; the study covers four countries in five geographic cells. The mixed cell was collected in two markets only and spans several unrelated trades. It is reported for completeness but is not a sector in the sense the other three columns are, and its row should not be read as a sector comparison.

Two selection biases are visible and both run the same way. Businesses appearing in professional registers are likely better established than those that do not, and the legal cell is drawn substantially from high-revenue firms. Both would raise measured readiness. Proportions here should therefore be read as generous to the sites, which strengthens rather than weakens the finding that a majority fall in the gap defined in section 7.

---

6 · FINDINGS: RETRIEVAL AND PRECONDITIONS

6.1 What was reached

Condition that returned markup	Sites	Share
Identified research user agent, HTTPS, host as listed	358	84%
Common browser user agent, same URL	4	1%
Browser user agent, host form flipped	7	2%
Browser user agent, plain HTTP	15	4%
No markup from any condition	44	10%

384 sites (90%) returned markup to at least one condition, 26 of them only after the first request failed. A single-request instrument would have recorded all 26 as unreachable.

The recoveries are not uniformly caused. Fourteen of the fifteen sites reached over plain HTTP had failed HTTPS with an invalid certificate — live businesses behind a broken certificate chain, a condition that also produces browser warnings for human visitors.

Cause of first-request failure	Sites
Request refused (HTTP 403)	21
Invalid TLS certificate	20
Server error (HTTP 5xx)	9
No response within timeout	8
Rate limited (HTTP 429)	5
TLS handshake failed	4
TLS or network failure	1
Not found (HTTP 404)	1
Bad request (HTTP 400)	1
Total	**70**

All 70 were re-requested at the identical URL changing nothing but the user-agent string. Four served the browser; 66 behaved identically to both. **User-agent discrimination is rare in this sample.** The observed refusals are predominantly TLS faults, filtering rules, and server failures that would refuse any client.

Four sites publish a robots.txt permitting every crawler and then return 403 to a client identifying as one. Stated policy and deployed behaviour disagree, and nothing in the site's own configuration would reveal this to its operator.

6.2 The seven preconditions

Precondition	Met	Measured	Share
Not blocking AI crawlers	365	382	96%
— permits explicitly in robots.txt	307	324	95%
— publishes no robots.txt	58	58	—
Indexable (no noindex)	387	400	97%
Content present without JavaScript	336	384	88%
H1 heading present	272	384	71%
Structured data present	266	384	69%
Title and meta description	293	400	73%
Valid llms.txt at domain root	75	382	20%

The crawler-access row is split because it combines two states. RFC 9309 treats absence of a robots.txt as absence of restriction, so a site with no file scores as permitting. Fifty-eight of the 365 published no file at all — not a decision to permit, but the absence of a decision.

Content without JavaScript is scored where the non-rendered document yields at least 100 words. The cut-point is not load-bearing: at 50 words the share is 88.7%, at 100 words 87.5%, at 150 words 85.7%, and only five sites fall between 50 and 100 words.

6.3 Distribution

Restricted to the 371 sites for which all seven properties were observable.

Score	0	1	2	3	4	5	6	7
Sites	1	5	26	16	36	106	135	46

Mean 5.15 of 7. Forty-six sites (12%) met all seven; 84 (23%) scored four or below.

6.4 By geography and sector

	n	Crawler access	Indexable	No-JS	H1	Schema	Title+meta	llms.txt
US East	102	98%	97%	87%	66%	70%	81%	**30%**
US West	78	94%	97%	90%	71%	72%	76%	**25%**
United Kingdom	78	94%	99%	90%	72%	72%	74%	**14%**
Canada	98	94%	98%	88%	75%	67%	71%	**16%**
Netherlands	72	97%	92%	83%	70%	64%	60%	**9%**

Sector	n	Crawler access	Indexable	No-JS	H1	Schema	Title+meta	llms.txt
Home services	130	95%	95%	80%	68%	69%	72%	**27%**
Legal	122	93%	97%	96%	72%	66%	75%	**8%**
Medical & dental	128	97%	100%	88%	71%	75%	75%	**24%**
Mixed small business	48	100%	91%	85%	74%	63%	67%	**18%**

Twenty-seven per cent of home-services businesses publish a valid llms.txt against 8% of law firms, and the legal cell is drawn substantially from the largest firms in each country. One reading consistent with the data is supply-side: independent trades run on managed platforms where a single vendor can deploy a file across an entire client base, while large firms run bespoke estates under change control. Nothing here tests that.

The geographic ordering is robust to how the mixed cell is assigned: under an alternative rule by country-code top-level domain the llms.txt column reads 31 / 25 / 14 / 14 / 9, unchanged in ordering.

6.5 Crawler blocking

Measured on 382 sites; the remainder returned no readable robots.txt to any condition and are reported as unmeasured rather than assumed open.

Measure	Sites	Share of measured
Blocking one or more AI crawlers by name	13	3%
Blanket disallow to all user agents	4	1%

Crawler	Sites blocking
Bytespider	10
CCBot	8
meta-externalagent	7
GPTBot	6
ClaudeBot	6
Google-Extended	6
Amazonbot	6
Applebot-Extended	6
ExaBot	2
PerplexityBot, Amzn-SearchBot, Diffbot, YouBot, TavilyBot, FirecrawlAgent	1 each

Declared blocking is not where the measured shortfall is concentrated. Ninety-five per cent of sites with a readable robots.txt permit every AI crawler by name. This differs sharply from the pattern Longpre et al. document among the high-traffic publisher domains underlying training corpora,[6] and the divergence is informative: restriction of AI crawlers appears to be a behaviour of large content owners, not of small businesses. Where owners here have made a decision, the ordering suggests it concerns training-data collection rather than answer-time retrieval, though the counts are too small to support that as a finding.

7 · THE CITATION GAP, OPERATIONALISED

The term is used loosely to mean that AI visibility differs from search visibility. That is a slogan. This section defines it as a computable partition of the sample.

7.1 Definition

Partition the seven preconditions by function (section 3). Then for each site:

- **Reached** — returned markup under at least one retrieval condition.
- **Access-eligible** — reached, *and* not blocking AI crawlers, *and* indexable. A retrieval client can request the page, is not technically disallowed by the site's robots policy, and receives a document.
- **Extraction-complete** — all four extraction preconditions observed and satisfied: readable text without JavaScript, an `h1`, structured data, and a title with meta description. The document meets the protocol's operational threshold for extraction readiness.

A note on the access criterion. Section 2.1 distinguishes indexing crawlers, collection crawlers, answer-time retrieval clients, and assistant fetchers as different mechanisms. The access criterion nonetheless treats a rule against any named AI crawler as blocking, which is the conservative reading: it counts a site as blocked on the broadest interpretation of the owner's expressed intent, rather than adjudicating which agents a given engine would use at answer time — an assignment the engines do not publish and which changes as products change.

The choice is empirically immaterial here. Of the 13 sites blocking a named AI crawler, 8 block at least one client oriented to answer-time retrieval and 5 block only collection or training crawlers. Recomputing the partition under the narrower criterion — counting only rules against retrieval-oriented clients — moves access-eligible sites from 349 to 353 and leaves the Citation Gap at 187, a rate of 53.0% against 53.6%. No result in this paper depends on the resolution.

A structural detail is worth recording: 6 of the 13 block an identical set of eight crawler tokens, and 4 more block a single token each. The identical blocks suggest a shared template or security-vendor default rather than 6 independent decisions, which is consistent with the broader finding that access outcomes in this sample are often not authored by the business.

> **The Citation Gap** is the set of sites that are access-eligible but do not satisfy all four extraction-readiness preconditions defined by this protocol.

These sites fall below the protocol's operational threshold for extraction readiness. The study does **not** claim that an individual engine would be unable to extract a claim from them. None of the four preconditions is established anywhere as a necessary condition for extraction by any specific system, and a page lacking all four may still contain readable, attributable prose. What the threshold provides is a consistent, externally verifiable criterion applied identically to every site in the sample — which is what makes the partition computable and the comparison across countries, sectors, and technology stacks meaningful.

The definition is deliberately conservative. It excludes unreached and blocked sites, which have a different and more visible problem. It sets aside llms.txt entirely, so the gap does not depend on an emerging convention. And it requires all four extraction preconditions to be *observed*, so no site enters the gap through missing data.

7.2 Size

Layer	Sites	Share of 428
L0 — no markup returned under any condition	44	10.3%
L1 — reached, but blocking or non-indexable	35	8.2%
L2 — access-eligible, not extraction-complete (Citation Gap)	**187**	**43.7%**
L3 — access-eligible and extraction-complete	162	37.9%

Of the 349 access-eligible sites, 187 (53.6%) fall in the Citation Gap. A slight majority of the sites an engine can reach and is permitted to use fall below the protocol's extraction-readiness threshold.

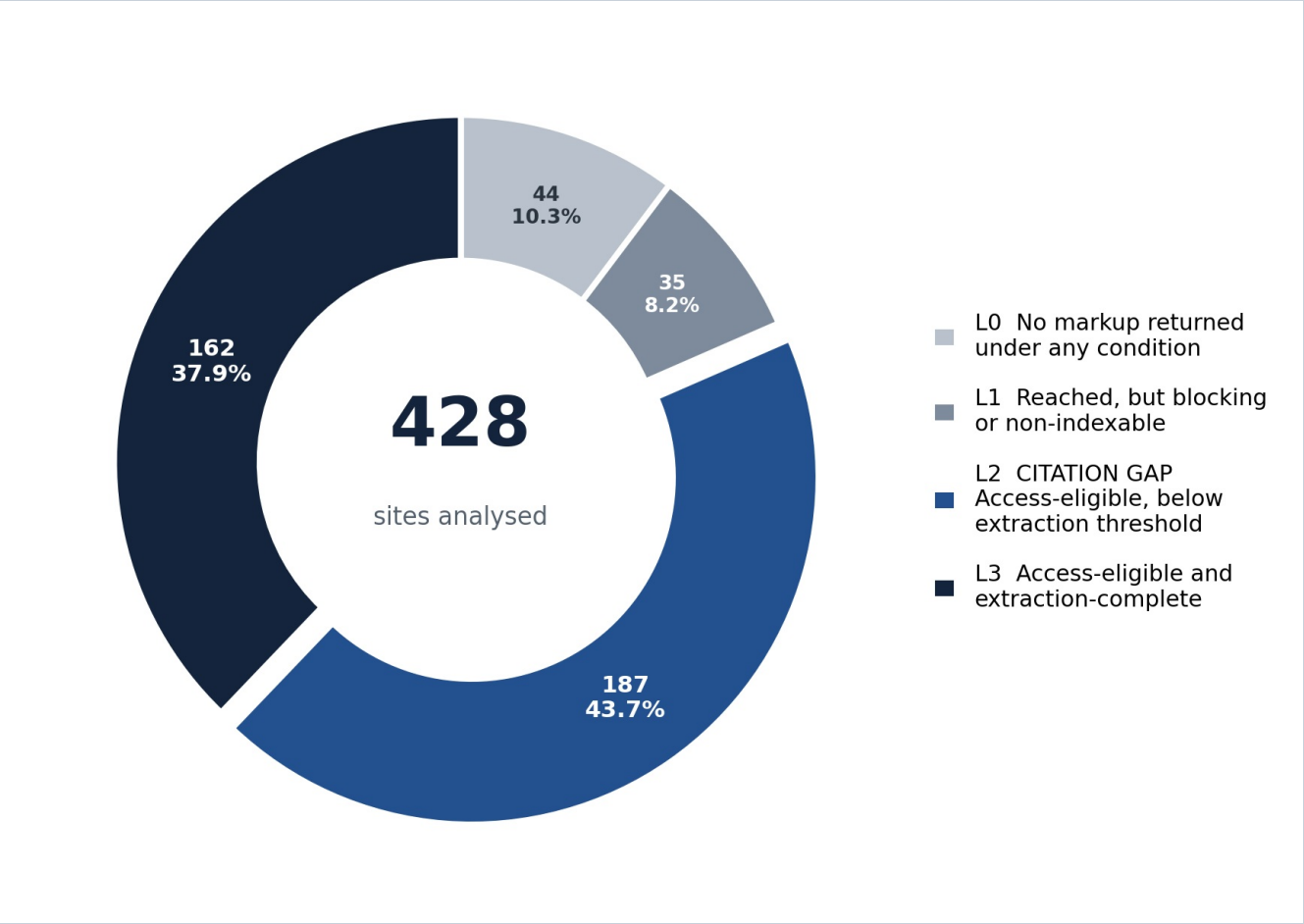

Figure 4. Partition of the 428 analysed sites. L0 and L1 are visible failures — the site returns nothing, declines, or de-indexes itself. L2, the Citation Gap, contains sites that resolve, respond, permit crawlers and index normally, but fall below the protocol's extraction-readiness threshold. Shares are of all 428 analysed sites; the 53.6% figure quoted in the text is L2 as a share of the 349 access-eligible sites (L2 + L3).

The layered form matters for interpretation. L0 and L1 are visible failures: the site is down, misconfigured, or has declined. L2 is invisible. These sites resolve, respond, permit, and index. They appear in search results. Nothing in their own analytics distinguishes them from L3.

7.3 Depth

Extraction preconditions failed	Sites	Share of gap
One	114	61%
Two	36	19%
Three	10	5%
All four	27	14%

Mean 1.73 of four. **Sixty-one per cent of gapped sites fail exactly one extraction precondition**, and the distribution of which one is concentrated:

Sites failing only this precondition	Count
H1 heading absent	45
Structured data absent	38
Title or meta description absent	31
Content without JavaScript absent	0

Not one site in the gap fails solely on non-JavaScript content. Where a site renders only in a browser, that condition arrives with others; where a site is a single precondition from the threshold, it is always a markup element rather than an architecture.

This is the paper's most consequential descriptive result, and it should be read carefully. It does not establish that adding an `h1` changes any engine's behaviour, nor that an `h1` is required for extraction. It establishes that under this operationalisation, the majority of the gap consists of sites separated from the threshold by one element of ordinary HTML. The gap measured here is therefore shallow rather than catastrophic — a distinction that matters both for interpretation and for what a follow-up study

would need to test.

7.4 Composition

	Access-eligible	In gap	Gap rate
US East	84	43	51.2%
US West	63	33	52.4%
United Kingdom	65	32	49.2%
Canada	80	43	53.8%
Netherlands	57	36	63.2%
Home services	99	47	47.5%
Legal	99	63	63.6%
Medical & dental	109	56	51.4%
Mixed small business	42	21	50.0%

Legal has the highest gap rate at 63.6% while also having the highest rate of content without JavaScript (96%). Law firms in this sample publish substantial server-rendered text and mark it up least: 66% carry structured data against 75% for medical and dental. The gap is not a content deficit in this cell; it is a markup deficit sitting on top of content.

7.5 The declaration layer does not close the gap

	With valid llms.txt
Gapped sites (L2)	27 of 187 — 14%
Extraction-complete sites (L3)	46 of 161 — 29%

Adoption is twice as high among sites already outside the gap. Eight sites publish a valid llms.txt while failing two or more extraction preconditions — a machine-readable index pointing at pages that fall well below the protocol's extraction-readiness threshold.

The observation is descriptive and its direction is not established. In this sample, llms.txt adoption is **associated with an already stronger technical baseline rather than explaining the measured gap**. Sites with more capable technical maintenance may adopt both, or the same vendor may supply both. What the data does not support is the inference that publishing the file moves a site out of the gap: the gap is defined without reference to it.

7.6 What separates the extremes

The 84 sites scoring four or below on all seven, against the remaining 287.

Precondition	Lowest 84	Remaining 287	Difference
Structured data present	13%	87%	**−74 pts**
Title and meta description	21%	88%	**−67 pts**
H1 heading present	31%	83%	**−52 pts**
Content present without JavaScript	49%	99%	**−50 pts**
Valid llms.txt at domain root	4%	25%	−21 pts
Indexable (no noindex)	88%	100%	−12 pts
Not blocking AI crawlers	90%	98%	−7 pts

Differences are computed on unrounded values and may differ by one point from the difference of the rounded columns. They are descriptive contrasts between two groups defined by the score itself, not effects.

Four extraction preconditions separate the extremes by fifty points or more; the declaration layer separates them by twenty-one. Both this and section 7.5 point the same way from different directions.

---

8 · MEASUREMENT VALIDITY OF THE EXTERNAL CHECK

Google's Lighthouse agentic-browsing category was requested for every site and returned a scored result for 402 of 428. Its accessibility-tree audit tests whether interactive elements carry names an automated client can resolve.

Result	Sites	Share of scored
Passes	137	34%
Fails	265	66%

Because a two-thirds failure rate is a strong claim, the instrument computed its own implementation of the same concept — enumerating interactive elements in the served document and testing whether each resolves an accessible name. On the 332 sites where both returned a verdict:

	Google: fail	Google: pass	Total
Local implementation: fail	119	43	162
Local implementation: pass	106	64	170
Total	225	107	332

Raw agreement is 55% — on a binary measure, close to chance. Google fails 68% of these sites; the local implementation fails 49%. By direct element count the median site resolves accessible names on 98% of its interactive elements, suggesting the Lighthouse audit is close to all-or-nothing at page level rather than proportional.

Neither implementation is thereby wrong. The divergence establishes that **66% is a property of Lighthouse's current implementation of an experimental category, not a property of the sites.** The defensible reading is that 66% of scored sites fail Google's current agentic accessibility-tree audit. Any longitudinal use of this measure must control for the Lighthouse version, which the dataset records per row.

This is why the Citation Gap in section 7 is built entirely on the seven direct preconditions and not on this audit. Each of those is an observation of the served document — a tag is present or it is not — with no scoring model between the document and the value.

---

9 · DISCUSSION

9.1 An argument, labelled as such

The following is an argument from the measurement, not a finding of it.

If retrieval-based answering grew on the demand side at the rates its operators report, one might expect the supply side to have adjusted. Section 7 finds a majority of access-eligible sites in the gap, most of them one markup element from the threshold — elements that have been recommended practice for a decade and cost nothing.

A mechanism consistent with this is the absence of feedback. A site that is not selected receives no signal: no rejection notice, no diagnostic in an analytics console, no record of answers in which another source was named. Conventional search at least returned a position. For the 44 sites here that returned nothing to any retrieval condition, there is not even a request to observe.

This explanation is untested. The data is a single cross-section and cannot distinguish absent feedback from absent attention, absent resources, or absent belief that the channel matters.

9.2 Where the constraint sits

The binding constraints in this sample are largely not editorial decisions. An invalid TLS certificate is a hosting fault. A managed bot-protection rule is a security product's default. A client-side-only architecture was chosen by whoever built the site, usually years earlier, on unrelated grounds. Twenty of 70 first-request failures were invalid certificates; four sites permit crawlers in writing and refuse them in practice.

For a small business this reframes the question. It is not what to publish, but whether anyone has checked what the site returns to a client that is not a browser — a question with a free answer, which Appendix A provides.

9.3 On llms.txt

Twenty per cent adoption is high enough to warrant scepticism and was tested. Every valid file was fingerprinted by heading

and body hash. No two sites returned a byte-identical file; headings name the individual business; three of 75 matched a generator-boilerplate pattern; sizes range from 590 bytes to 177 kilobytes, median 4.8 kilobytes. The adoption is genuine.

Its relationship to the gap is the finding. Sections 7.5 and 7.6 show adoption concentrated outside the gap and separating the extremes by a fifth of what structured data does. A convention no engine is documented as requiring is not where the measured shortfall sits in this sample.

9.4 What this cannot support

No causal claim. Nothing observes engine behaviour. Sites in the gap fall below this protocol's operational threshold for extraction readiness; whether that corresponds to any difference in retrieval, selection, or commercial outcome is not addressed, and no precondition used here is established as necessary for extraction by any engine. The inference from eligibility to consequence is an argument, set out above, not a finding of this dataset.

---

10 · LIMITATIONS

- The sample is designed to exercise the protocol across heterogeneous sites, not to estimate prevalence. No proportion here is a population estimate and no significance testing is reported.
- Two visible selection biases (professional-register listing, large-firm legal cell) both raise measured readiness, so reported proportions are generous to the sites.
- Denominators vary by property and by site. Every table states its own.
- One homepage request per site may not represent the site as a whole. The Citation Gap is a homepage-level measure.
- Measurements describe 16–18 August 2026. No claim about change over time is supported.
- Nothing measures whether any generative system retrieves, cites, or recommends any site.
- The four extraction preconditions constitute an operational threshold defined by this protocol. None is established as a necessary condition for extraction by any specific engine, and a site below the threshold may still contain readable, attributable text.
- The four extraction-readiness preconditions are an operational threshold defined by this protocol. None is a documented necessary condition for extraction by any engine, and the Citation Gap measures position relative to that threshold only.
- The agentic accessibility measure is implementation-specific and experimental (section 8), and is excluded from the Citation Gap for that reason.
- Indexability is read from the robots meta tag only; an `X-Robots-Tag` response header would not be detected.
- Sector labels were assigned at collection from the source directory, not inferred from page content, and the mixed cell is not a sector.
- User-agent attribution in first-party crawler observations is unverified and should be read as an upper bound.
- Three sites were lost to instrument failure rather than site behaviour; a small number of rows lost later retrieval conditions to a per-invocation subrequest ceiling, recorded in the row.

---

11 · CONCLUSION

Existing measurement of the web observes the rendered page, the stated policy, the crawled corpus, or content already retrieved. This paper specifies and applies a protocol that observes what a non-rendering identified client receives, and uses it to give *Citation Gap* an operational definition: sites an engine can reach and is permitted to use, but which do not satisfy all four extraction-readiness preconditions the protocol specifies.

In this sample of 428 business websites, 53.6% of access-eligible sites fall below that threshold, and 61% of those fail exactly one of the four extraction preconditions — most often a heading, structured data, or a meta description. Not one fails solely on rendering architecture. The emerging convention designed for this problem is adopted at twice the rate outside the gap as inside it, and is therefore associated with an already stronger technical baseline rather than with crossing the threshold.

The gap so defined is invisible from inside the affected business. These sites resolve, respond, permit crawlers, and index. What the protocol adds is a way to see the difference, cheaply and from outside, and Appendix A reduces it to a free forty-five-minute procedure.

Whether closing the gap changes any engine's behaviour is not established here, and would require observation of engine

outputs over time against sites whose preconditions changed. That is the natural next study, and this protocol is designed to supply its independent variable.

REFERENCES

[1] Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. (2024). GEO: Generative Engine Optimization. *KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5–16. https://doi.org/10.1145/3637528.3671900

[2] Chen, M., Wang, X., Chen, K., and Koudas, N. (2025). Generative Engine Optimization: How to Dominate AI Search. arXiv:2509.08919.

[3] Kumar, A., and Lakkaraju, H. (2024). Manipulating Large Language Models to Increase Product Visibility. arXiv:2404.07981.

[4] Wan, A., Wallace, E., and Klein, D. (2024). What Evidence Do Language Models Find Convincing? *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 7468–7484.

[5] WebAIM (2026). *The WebAIM Million: The 2026 Report on the Accessibility of the Top 1,000,000 Home Pages.* Utah State University, February 2026 analysis, published 30 March 2026. https://webaim.org/projects/million/

[6] Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., et al. (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons. *Advances in Neural Information Processing Systems 37* (NeurIPS 2024 Datasets and Benchmarks). arXiv:2407.14933.

[7] Sun, Y., Zhuang, Z., and Giles, C. L. (2007). A Large-Scale Study of robots.txt. *Proceedings of the 16th International Conference on World Wide Web*, 1123–1124.

[8] Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2199–2208.

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33*.

[10] Koster, M., Illyes, G., Zeller, H., and Sassman, L. (2022). *Robots Exclusion Protocol.* RFC 9309, Internet Engineering Task Force. https://www.rfc-editor.org/rfc/rfc9309

[11] Howard, J. (2024). *The /llms.txt file.* Answer.AI. https://llmstxt.org

[12] Adobe Digital Insights (2025). *AI Is Changing How Consumers Engage with Websites.* February 2025. Source for AI-referred traffic growth, July 2024 to February 2025, US market.

[13] W3C (2023). *Web Content Accessibility Guidelines (WCAG) 2.2.* W3C Recommendation.

[14] Google (2026). *AI Overviews* — reported reach of two billion monthly users. Company statement.

[15] Alphabet Inc. (2026). *Second Quarter 2026 Results* — Gemini application monthly active users.

[16] Anthropic. *Web search tool documentation* — citations enabled by default, returning source URL, title, and snippet for every result. https://docs.claude.com

[17] Chrome Developers (2026). *Agentic browsing audits.* Lighthouse 13.3, May 2026. https://developer.chrome.com — category documented as experimental and under development; not a confirmed ranking factor.

[18] Google. *PageSpeed Insights API v5.* https://developers.google.com/speed/docs/insights/v5/get-started

[19] Vikas (2026). *Generative Engine Optimization for Website Owners.* vSourceCode. https://vsourcecode.com/reports/book/geo-for-website-owners

[20] Vikas (2026). *The Missing Half of YouTube SEO.* vSourceCode. https://vsourcecode.com/reports/book/the-missing-half-of-youtube-seo

Sourcing discipline: only figures disclosed by the measuring party are cited as fact. Third-party panel estimates are labelled as estimates. Statistics-aggregator secondary sources are not used.

Appendix A — Agent-Readiness and Observed Citation Self-Assessment

A printable worksheet. No tools, no signup, no cost.

Complete in one sitting, roughly 45 minutes. Repeat at day 30 and day 60 and compare. The value is in the second run, not the first — a single score tells you where you are, two runs tell you whether anything you did worked.

Business _____ **Website** _____

Primary service _____ **Primary service area** _____

Assessor _____ **Date** _____ **Round** ☐ Baseline ☐ Day 30 ☐ Day 60

What this worksheet measures, and what it does not

An agent-readiness audit measures whether an agent can technically understand and interact with a website. It does not measure whether a generative search engine will retrieve, select, cite, recommend, or rely upon that website.

Six limits apply throughout, to the published research as much as to your own run:

- Agentic browsing standards remain experimental.
- User-agent identification is imperfect.
- A single-domain dataset cannot represent the entire web.
- Generative systems change rapidly.
- Citation and recommendation behaviour is not deterministic.
- The protocol measures technical readiness, not business outcomes.

Part 1 — Machine readability

Google’s Lighthouse Agentic Browsing category, reported in PageSpeed Insights as a pass ratio rather than a 0–100 score.

How to run it. Open `pagespeed.web.dev`, enter your homepage URL, select Mobile, press Analyze. Repeat for Desktop and for your most important service page.

> The category is experimental and marked under development. It is not a confirmed ranking factor. It is a readability diagnostic.

Page	Device	Perf.	Access.	Best Pr.	SEO	Agentic
Homepage	Mobile	____	____	____	____	___ / ___
Homepage	Desktop	____	____	____	____	___ / ___
Key service page	Mobile	____	____	____	____	___ / ___

Verification URL (press “Copy Link” in PSI and paste — this makes your result independently checkable):

☐ **Accessibility tree is well-formed.** Agents read the accessibility tree — the map of roles, names and states — as a screen reader does. If a button is an unlabelled `div` with a click handler, software cannot identify it as a button. *The practical question: do your buttons, links and form fields have real labels?*

> This audit is implementation-specific. Independent implementations of the same concept disagree on the same page (section 7 of this paper). Treat the result as one tool’s reading, and record the Lighthouse version shown in your report.

Failing items: _____

☐ **Layout is stable (Cumulative Layout Shift).** Google’s guidance is explicit: agents that take screenshots are confused by shifting layouts, and an agent moves faster than a person who would simply wait.

Your CLS: _____ (Good ≤ 0.1 · Needs work 0.1–0.25 · Poor > 0.25)

☐ **A valid llms.txt exists at the domain root.**

Present at `yourdomain.com/llms.txt`? ☐ Yes ☐ No · Has an H1 heading? ☐ Yes ☐ No · Contains links? ☐ Yes ☐ No

> Many sites return the homepage for any missing path. If that address shows your normal web page rather than plain text, you do not have the file — you have a fallback.

Part 1 result: _____ / 3

Part 2 — Crawl access

A page that cannot be retrieved by an engine cannot be directly cited by that engine. This part costs nothing and is the most common silent failure.

How to check. Type `yourdomain.com/robots.txt` into a browser. If nothing loads, you have no robots.txt — usually fine, but note it.

Question	Yes	No
Is there a blanket `Disallow: /` under `User-agent: *`?	☐	☐
Is any AI crawler named and disallowed?	☐	☐
Is your sitemap listed?	☐	☐
Do your key service pages load without a login?	☐	☐

Crawlers named and blocked in your file: _____

> Crawler names are matched exactly. `Applebot-Extended` and `Applebot` are different agents, and a rule written for one does not govern the other.

Server-side blocking. Firewall and bot-protection products can refuse AI crawlers even when robots.txt permits them. Check your CDN or security dashboard for managed rules blocking AI bots, and for challenge modes issuing a JavaScript test — most crawlers do not execute JavaScript and fail silently.

Bot protection in use: _____ AI-bot rule enabled? ☐ Yes ☐ No ☐ Don’t know

> Your robots.txt and your edge configuration are set in different places and can contradict each other. A permissive robots.txt does not prove a crawler is being served. Nothing in your own site’s settings will show you this; it is visible only in your CDN or firewall logs.

Part 2 result: ☐ Open ☐ Partially blocked ☐ Blocked

Part 3 — Observed Citation Count

Parts 1 and 2 measure whether a machine *can* read you. This part records whether any engine named you in a given set of runs. **Passing Parts 1 and 2 makes you eligible. It does not make you cited.**

Build your question set. Write ten questions a stranger would type — not ones containing your business name.

> The single most common error is testing your own name. That tests whether the engine knows you exist. It does not test whether the engine recommends you to somebody who has never heard of you, which is the only question that matters.

1. best [SERVICE] in [CITY] · 2. how much does [SERVICE] cost in [CITY] · 3. how do I choose a good [PROFESSIONAL] in [CITY] · 4. [CUSTOMER’S PROBLEM] — what should I do · 5. is it worth paying for [SERVICE] · 6. [SERVICE] near me — what should I look for · 7. what questions should I ask before hiring a [PROFESSIONAL] · 8. how long does [PROCESS] take in [REGION] · 9. [SERVICE A] or [SERVICE B] — which do I need · 10. what does [SERVICE] involve

Run and record. Use a logged-out or private window so your own history does not bias the answer.

Engines: ☐ ChatGPT ☐ Perplexity ☐ Google AI Overviews ☐ Gemini ☐ Copilot ☐ Claude ☐ Other _____

#	Engine 1	Engine 2	Engine 3	Which source was named instead
1	☐	☐	☐	_____
2	☐	☐	☐	_____
3	☐	☐	☐	_____
4	☐	☐	☐	_____
5	☐	☐	☐	_____
6	☐	☐	☐	_____
7	☐	☐	☐	_____
8	☐	☐	☐	_____
9	☐	☐	☐	_____
10	☐	☐	☐	_____

Observed Citation Count: _____ out of 30 runs

The right-hand column is the most useful thing on this page. Those are the sources these engines selected for your topic in your area, on these runs. Study what they have that you do not.

Most frequently named source: _____ Appeared in _____ of 30 runs.

> Engines are non-deterministic. The same question can return different sources on different runs. Treat a single run as a sample, not a measurement, and compare like-for-like across rounds.

Scoring

	Result
Part 1 — Machine readability	_____ / 3
Part 2 — Crawl access	☐ Open ☐ Partial ☐ Blocked
Part 3 — Observed Citation Count	_____ / 30

Pattern	What it means	Do this first
Part 2 blocked	Nothing downstream can be assessed yet	Fix access. Everything else is wasted effort until crawlers can reach you.
Part 2 open, Part 1 low	Machines reach you but struggle to parse you	Labels and layout stability. Both are ordinary accessibility work.
Parts 1 and 2 good, Part 3 zero	Technical eligibility alone has not resulted in observed selection	Possible explanations include content, authority, relevance, source coverage, query formulation, or engine-specific selection. Study the right-hand column before assuming which applies.
Part 3 above zero	At least one engine named you in at least one run	Widen the question set and track which categories you lose.

Round	Date	Part 1	Part 2	Part 3
Baseline	_____	___/3	_____	___/30
Day 30	_____	___/3	_____	___/30
Day 60	_____	___/3	_____	___/30

Keep the question set identical between rounds. Changing the questions changes the measurement, and you lose the only thing that made the second round worth doing.

Parts 1 and 2 use publicly documented criteria published by Google. Part 3 is a self-administered observation protocol; its results are non-deterministic and should be treated as indicative rather than as a measurement. Nothing in this worksheet establishes that any change made to a website caused any change in citation behaviour.

Appendix B — Dataset and availability

Unit of observation. One row per audited domain, 447 rows, of which 428 are in the analytical set.

Recorded per row. Retrieval outcome under each of the four conditions and the full attempt ladder; final HTTP status and origin; parked-domain determination and the layer that detected it; robots.txt status, parsed groups, sitemap declarations, per-crawler block determinations under both strict and permissive token matching; llms.txt status, validity components, byte length, heading line, and body hash; document-derived properties (word count, headings, title, meta description, canonical, JSON-LD types, microdata, language, robots meta, landmarks, image and form labelling); PageSpeed Insights category scores and Core Web Vitals; the Lighthouse agentic-browsing audit results and version; a local replication of the accessibility-tree audit with element counts; and, for each of the seven preconditions, both the value and the observer that supplied it.

Site identifiers are withheld. This study reports on identifiable third parties who did not consent to participation and could be adversely affected by attribution. Every finding here is distributional; naming the sites that scored poorly would expose identifiable small businesses to reputational cost for no research benefit. Withholding identifiers in these circumstances is standard practice in web measurement, not a limitation of the work.

Availability. The anonymised record-level dataset — one row per site with a synthetic identifier, country, sector, and every measured value, containing no domain names or URLs — accompanies this paper. The identified dataset, including per-domain screening records and PageSpeed verification links for every audit, is retained in full and available to reviewers, editors, and replicating researchers on written request to the author.

Reproducibility. Every measure is deterministic. Every figure in this paper derives from a single classification applied to one locked dataset, so no table can diverge from another. All tables were regenerated from that file and independently recomputed against it.

Appendix C — Pre-audit screening record

The source list carried 484 unique domains. Thirty-seven did not resolve at screening and were not queued. They were re-measured under the full protocol on 18 August 2026 and all 37 returned `dns_no_address`: no address record at the apex or `www` host, under either protocol. They are recorded here rather than dropped silently, because a 7.6 per cent rate of non-resolution in a frame drawn from live public directories is itself an observation about the small-business web.

	Home services	Legal	Medical & dental	Mixed	Total
US East	1	0	1	2	**4**
US West	3	0	2	0	**5**
United Kingdom	1	5	4	0	**10**
Canada	2	0	4	0	**6**
Netherlands	2	4	6	0	**12**
Total	**9**	**9**	**17**	**2**	**37**

Expressed as a rate against each sector's listed domains, non-resolution is roughly twice as common among medical and dental practices as among the other sectors:

Sector	Non-resolving	Listed	Rate	Also parked	Combined
Medical & dental	17	150	11.3%	5	**14.7%**
Home services	9	149	6.0%	10	**12.8%**
Legal	9	135	6.7%	1	**7.4%**
Mixed small business	2	50	4.0%	0	**4.0%**

The combined column counts domains listed in a public register or directory that have no working website at the listed address — either no DNS record, or a domain-parking page. On that measure roughly one in seven listed medical and dental practices, and one in eight home-services businesses, is unreachable at its published address, against one in fourteen law firms.

This study cannot distinguish among the possible causes, which include practice closure, merger or rebranding, lapse of a domain registration, migration to a social or marketplace profile without a website, and staleness in the register itself. The measure is domain resolution, not business survival, and no inference about closure rates should be drawn from it. It is reported because it bears directly on the construction of any sample drawn from directory sources: a frame compiled from professional registers carries a non-trivial and sector-dependent share of entries that no longer resolve, and a study that silently drops them understates both the attrition and its unevenness.

Two further domains initially screened out proved to be transcription errors in the source list rather than dead domains. Both were re-audited under identical conditions and are included in the analytical set: one returned HTTP 522 under all four retrieval conditions and is retained as a site returning no readable markup, the other returned 200 and scored six of seven.

Per-domain status is retained in the identified dataset and available on the terms in Appendix B.

Appendix D — Aggregate findings

This appendix consolidates the measured results in a single reference table set. All figures are for the 428 analysed sites; denominators differ by measure and are stated in each row.

D.1 Retrieval

Condition returning markup	Sites
Identified research UA, HTTPS, host as listed	358
Browser UA, same URL	4
Browser UA, host form flipped	7
Browser UA, plain HTTP	15
No markup from any condition	44

Reached by at least one condition: 384 of 428 (90%). Reached only after first-request failure: 26. First-request failures: 70. User-agent differential: 4 of 70. Sites publishing permissive robots.txt and refusing an identified crawler: 4.

D.2 Preconditions

Precondition	Met	Measured	Share
Not blocking AI crawlers	365	382	96%
Indexable (no noindex)	387	400	97%
Content without JavaScript	336	384	88%
H1 heading present	272	384	71%
Structured data present	266	384	69%
Title and meta description	293	400	73%
Valid llms.txt at root	75	382	20%

Sixteen sites carry a Lighthouse-sourced value for indexability and title/meta-description, accounting for the denominator of 400 on those two rows. Twenty-nine sites had Lighthouse values withheld for offsite redirection.

D.3 Score distribution (371 complete cases)

0	1	2	3	4	5	6	7
1	5	26	16	36	106	135	46

Mean 5.15. Scoring ≤4: 84 (23%). Scoring 7: 46 (12%).

D.4 llms.txt verification

Check	Result
Valid files found	75
Files sharing a byte-identical hash with another site	0
Files flagged as generator boilerplate	3
Median file size	4,762 bytes
Range	590 – 177,155 bytes

D.5 Agent accessibility

Scored 402 of 428. Pass 137 (34%), fail 265 (66%). Local replication agreement with Lighthouse: 55.1% on 332 sites where both returned a verdict.

D.6 Document control

Sites analysed	428
Source list, unique domains	484
Measurement window	16–18 August 2026
Retrieval conditions per domain	Up to 4, all recorded
Preconditions observed per site	7, each with its supplying observer
Second observer	Google PageSpeed Insights / Lighthouse 13.3
Numerical verification	89 reported quantities recomputed from the locked dataset; 89 agree
Identifiers	Withheld; see Appendix B

Every quantity in this paper, including every denominator, was regenerated from a single locked dataset and independently recomputed against it before release. Where a figure in this paper disagrees with any earlier draft, working note, or preliminary release, the figure here supersedes it.

Requests for the identified dataset, the per-domain screening record, or the PageSpeed verification links should be addressed to the author.

Measuring the Citation Gap. Vikas, vSourceCode, August 2026. n = 428 sites analysed, 16–18 August 2026.