AIO MAPPER

ORIGINAL RESEARCH / SEPTEMBER 2026

The AI visibility starting line

What 1,000 popular domains reveal about crawler access, content delivery, and the limits of conventional SEO signals

AIO Mapper Research

The market is debating advanced AI optimization while many domains still have not cleared the first requirement: delivering enough content for a crawler to work with.

One request per homepage. No JavaScript rendering. Every outcome retained.

WHAT LEADERS SHOULD KNOW

Executive Summary

Delivery is the first constraint.

Of 576 domains that returned HTML, 100 supplied fewer than 50 body words. At the applied threshold, only 476 of all 1,000 sampled domains reached the primary score analysis.

Failure is not one category.

The 499 domains outside the raw *ok* result included DNS failures, bot and WAF blocks, robots refusals, timeouts, and thin HTML. Treating them as one verdict would hide the operational cause.

Popularity did not protect delivery.

Among returned HTML, the top-1,000 rank band had the highest share below 50 words: 22.1%, versus 12.7% in the 100,000-to-1-million band.

SEO and AI readiness overlap only partly.

Across 476 qualifying pages, Spearman's rho was 0.264. That is a real but weak relationship, equivalent to roughly 7% shared variance.

Familiar signals are not enough.

JSON-LD and H1 presence aligned much more strongly with the SEO score than with the AI-readiness score. They remain useful, but they do not diagnose access or served-content failures.

Industry cuts require a new data layer.

The source list includes both websites and infrastructure domains, and the crawl did not record industry. This report uses delivery archetypes for the primary segmentation and proposes a pre-classified sector design for the next wave.

A DELIVERY MAP, NOT A PASS-FAIL SCORE

The starting line has seven distinct outcomes

The headline result becomes more useful when the sample is separated by what actually happened at retrieval. Nearly half delivered 50 or more words. The other half split across six operationally different states.

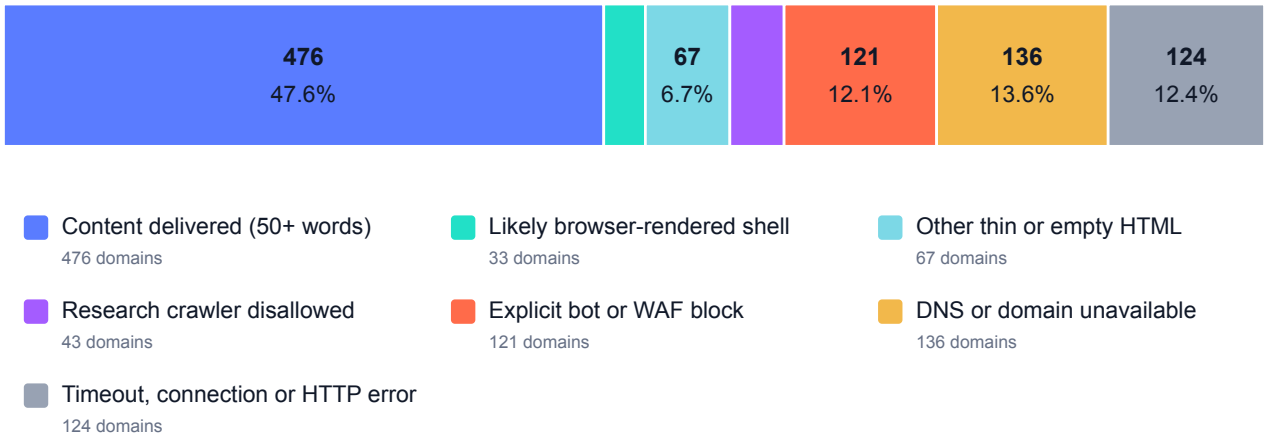

Exhibit 1. Mutually exclusive delivery archetypes across all 1,000 domains. "Likely browser-rendered shell" is based on the recorded client-rendering flag within thin HTML. Source: AIO Mapper study dataset.

> **So what:** A single "AI visibility failed" label cannot tell a team what to fix. DNS ownership, security policy, rendering architecture, robots controls, and content publishing each have different owners.

THE CAUSES POINT TO DIFFERENT OWNERS

A third of fetch failures were explicit bot or WAF blocks

The 381 fetch failures were not interchangeable. DNS or unavailable domains accounted for 136. Another 121 were explicitly attributed by the crawler to bot protection or a web application firewall.

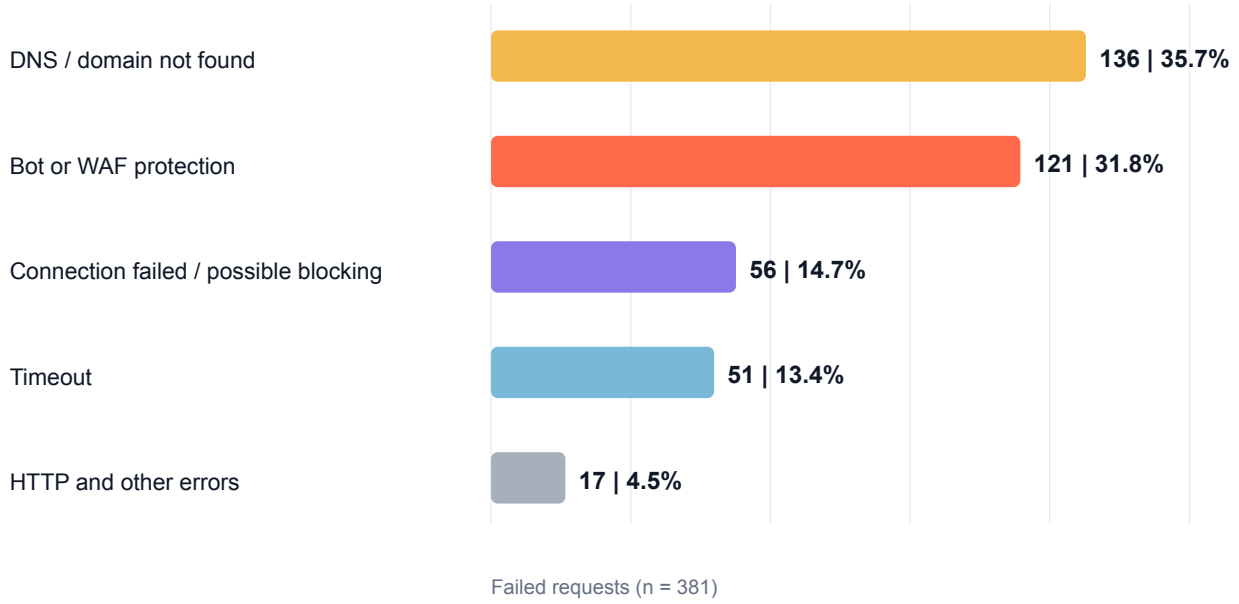

Exhibit 2. Recorded failure reasons among 381 fetch failures. The 56 connection failures may include blocking, but the message is ambiguous, so they remain separate. Source: AIO Mapper study dataset.

So what: The cleanest business opportunity is not to call every failure a website problem. Separate intentional policy from accidental blocking and dead or non-web domains before assigning remediation.

SENSITIVITY ANALYSIS

The finding survives stricter content thresholds

The 50-word rule was deliberately permissive. Raising it to 100 or 200 words increases the below-threshold share, but does not change the basic conclusion: a material portion of returned HTML contained little body text.

17.4%
below 50 words
95% CI 14.5%-20.7%

19.8%
below 100 words
95% CI 16.7%-23.2%

24.7%
below 200 words
95% CI 21.3%-28.3%

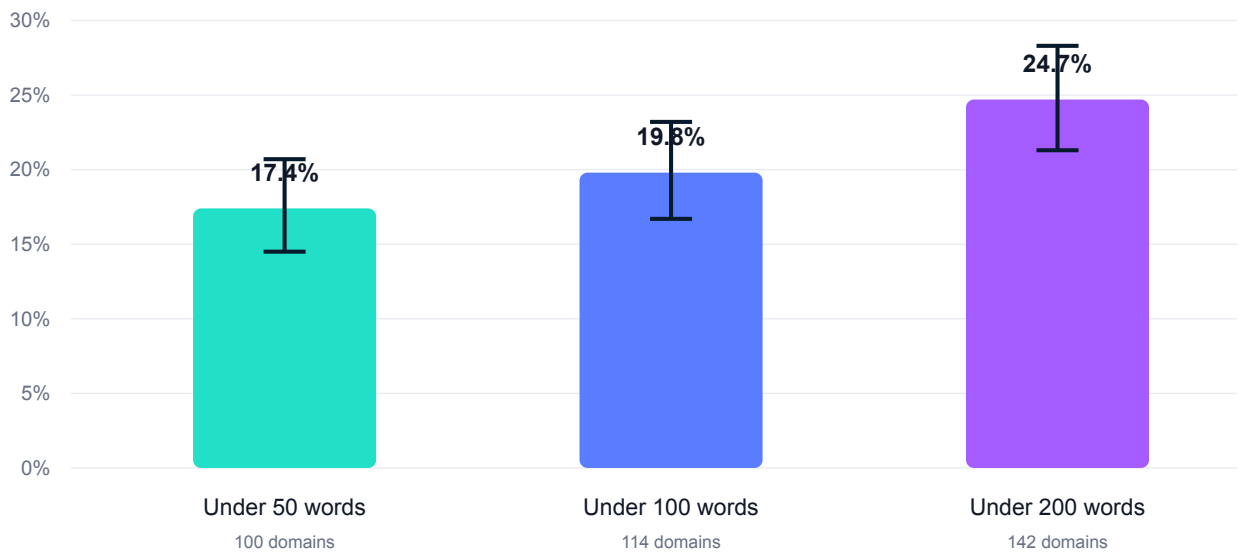

Exhibit 3. Threshold sensitivity among 576 HTML responses. Error bars show 95% Wilson intervals. Source: AIO Mapper study dataset.

So what: The story does not depend on selecting one unusually strict definition of content. The 50-word cut is the most conservative version of the result.

RANK-BAND ANALYSIS

Popularity did not eliminate delivery problems

HTML return rates ranged from 54.4% to 61.2% across the four equal rank bands. More surprisingly, the highest-ranked group had the largest share of thin HTML responses.

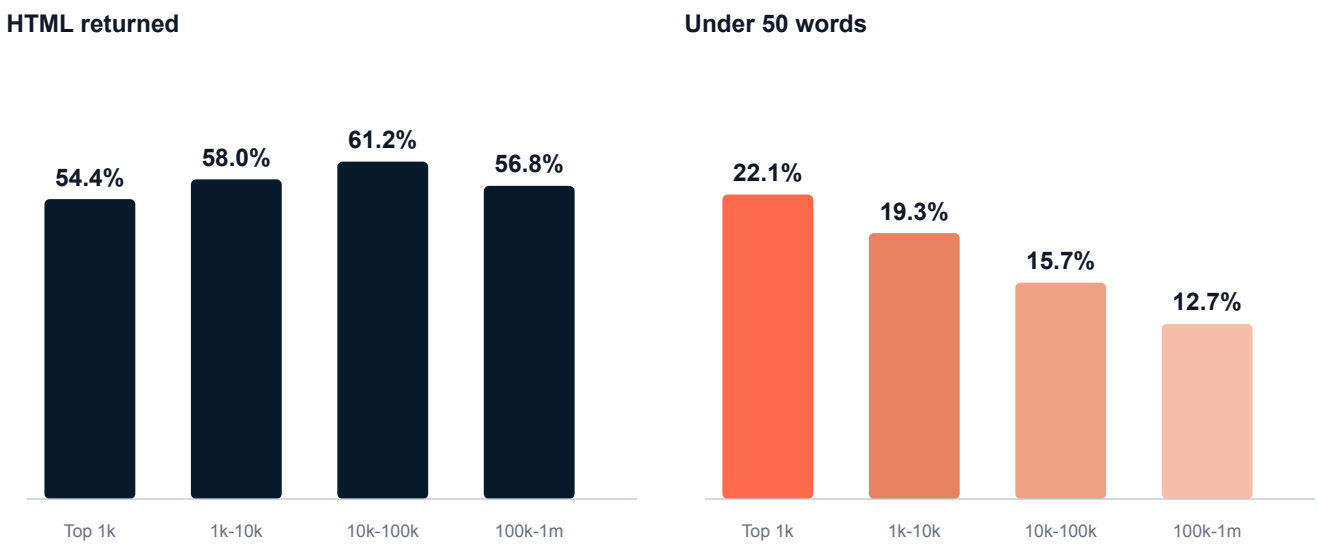

Exhibit 4. Each rank band contains 250 sampled domains. The right panel uses only domains that returned HTML as its denominator. Source: AIO Mapper study dataset.

Prominence can bring more sophisticated delivery stacks, but sophistication does not guarantee that a non-rendering crawler receives the useful page body.

This pattern is descriptive, not causal. The top band also contains widely used infrastructure and service domains, and one request cannot distinguish stable policy from a temporary response.

THE MEASURES ARE RELATED, NOT INTERCHANGEABLE

Traditional SEO explains only a small part of readiness

The two scores moved in the same direction, but not closely. Pages with similar SEO scores occupied a wide range of AI-readiness outcomes.

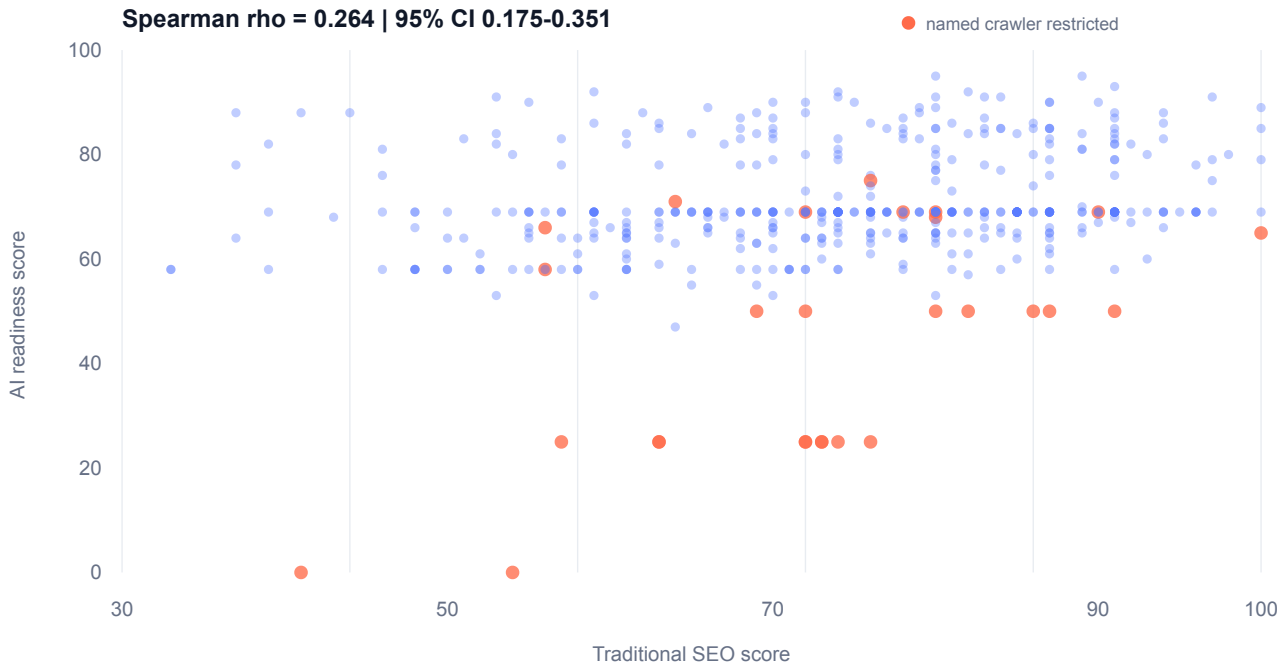

Exhibit 5. Qualifying content pages only (n = 476). Spearman's rho = 0.264, bootstrap 95% CI 0.175-0.351. Source: AIO Mapper study dataset.

So what: Conventional SEO remains relevant to source quality and discoverability, but it is not a substitute for testing the response seen by the crawler. Google likewise describes crawlability and technical eligibility as foundations for its generative search features.[6]

PERMISSIONS ARE PART OF DISTRIBUTION

Crawler access can dominate the readiness score

Among qualifying pages, 28 restricted at least one named answer-time AI crawler in the recorded crawlability component. Their average SEO score was close to the fully allowed group; their average AI-readiness score was 26.3 points lower.

Exhibit 6. Mean scores among 476 qualifying pages. The AI-readiness score includes crawler access by design, so this is an instrument-behavior comparison, not an independent causal estimate. Source: AIO Mapper study dataset.

So what: Robots policy has become a distribution decision. OpenAI states that blocking OAI-SearchBot prevents content from being included in ChatGPT summaries and snippets.[5] Anthropic separates training, search, and user-directed retrieval into different robots.[7]

A CONCRETE EXAMPLE FROM THE DATASET

One row shows why browser visibility is not crawler visibility

In the point-in-time research fetch, **nytimes.com** looked healthy by several conventional measures. The response was substantial and structured. Its recorded crawler permissions told a different story.

nytimes.com

1.1 MB	1,112	72	25
HTML response	body words	SEO score	AI readiness score

One H1. JSON-LD present. AI bot crawlability component: 25.

The page did not lack content. It did not lack structure. The divergence appeared at the permissions layer.

This does not show whether the New York Times appears in AI answers. It demonstrates the narrower claim this study can support: browser-visible content, familiar SEO signals, and crawler-level access are not the same thing.

THIN HTML HAS MORE THAN ONE CAUSE

Client rendering explains nearly half of unreadable responses - not all

Of the 75 responses the crawler marked unreadable, 33 showed enough scaffolding to suggest that client-side rendering may explain the missing body text. That is 44.0% of the subgroup.

75
responses marked unreadable

33
likely client-rendered

44.0%
95% CI 33.3%-55.3%

33
likely browser-rendered shell

42
other thin or empty HTML

So what: A rendered screenshot can look complete while the initial response remains empty. But JavaScript is not the only explanation; consent gates, localization, service endpoints, and transient responses can create the same symptom.

FAMILIAR SIGNALS IN CONTEXT

Schema and headings help - but they do not clear the starting line

Among qualifying pages, JSON-LD and an H1 were associated with higher scores. The differences were much larger for the SEO score than for AI readiness.

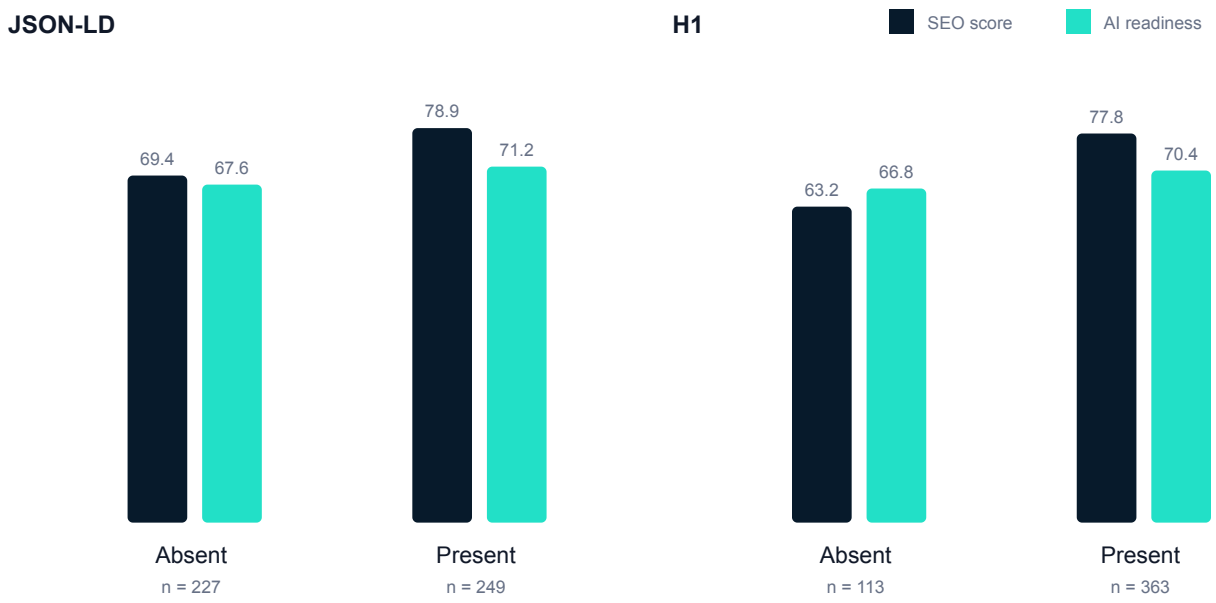

Exhibit 7. Mean scores by signal presence among 476 qualifying pages. These are descriptive associations and do not isolate causal effects. Source: AIO Mapper study dataset.

JSON-LD: presence aligned with a 9.5-point SEO difference and a 3.6-point AI-readiness difference.

H1: presence aligned with a 14.6-point SEO difference and a 3.6-point AI-readiness difference.

> **So what:** Structure is useful after access and delivery are working. Google explicitly cautions against overfocusing on structured data for generative search.[6]

RESPONSE WEIGHT IS WEAKLY RELATED TO READINESS

A heavy response is not automatically an unreadable response

The median returned response weighed 140 KB and contained 618 body words. Among qualifying pages, the relationship between response size and AI readiness was almost flat.

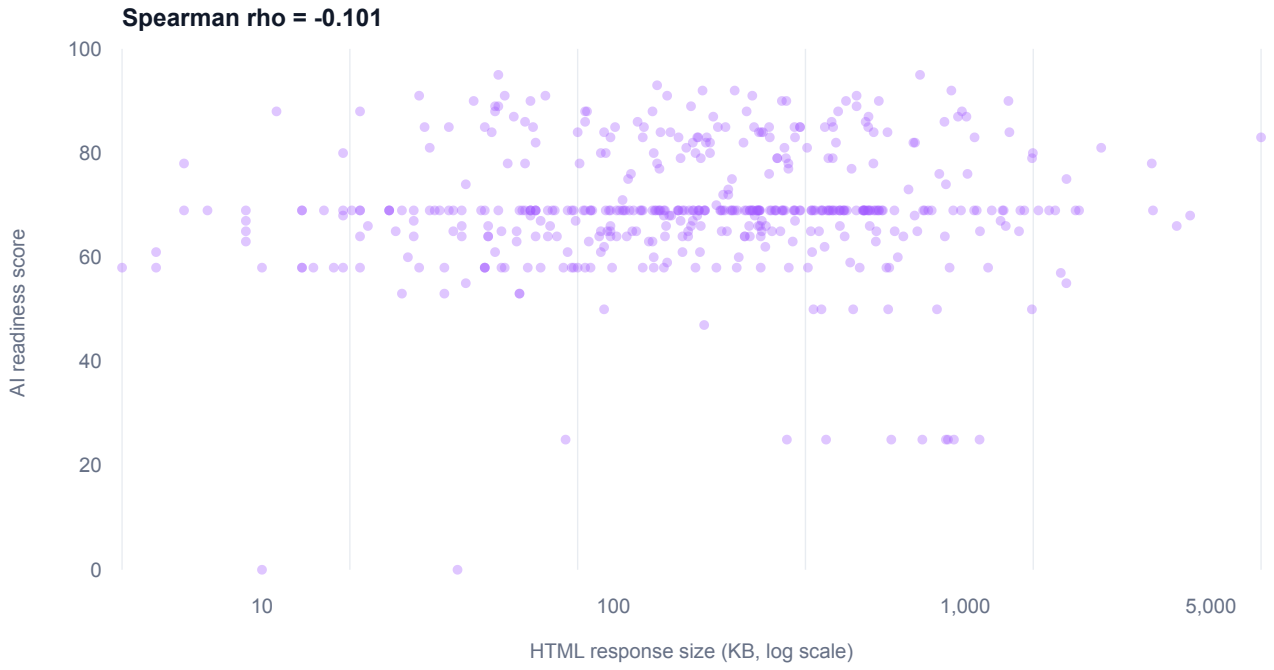

Exhibit 8. HTML response size and AI-readiness score among 476 qualifying pages. The x-axis is logarithmic. Source: AIO Mapper study dataset.

So what: "The page is too heavy" is not a sufficient diagnosis. What matters first is whether the useful text is present in the response, accessible under the relevant policy, and separable from the delivery shell.

A FIVE-STEP OPERATING MODEL

Treat AI visibility as a cross-functional delivery test

The evidence points to a practical order of operations. Teams should diagnose the earliest failed checkpoint before investing in downstream optimization.

1. **Request the page as the crawler you care about.**
   Record status, redirects, response body, timing, and any challenge page.

2. **Confirm that the main content is in the initial HTML.**
   Compare raw response text with the rendered browser view.

3. **Audit purpose-specific permissions.**
   Separate training crawlers from search and user-directed retrieval rather than applying one blanket rule.

4. **Check extractability.**
   Verify that coherent text, headings, entities, and evidence can be separated from navigation and interface chrome.

5. **Optimize only after delivery is reliable.**
   Then improve structure, original evidence, schema, internal context, and answer clarity.

AI visibility begins as an infrastructure and publishing question. It becomes a content optimization question only after the page reaches the machine.

RESEARCH AGENDA

The next wave should benchmark sectors and page templates

A follow-up study can answer the industry question cleanly if classification is designed before collection rather than inferred afterward.

Sampling

- Pre-classify domains with an external industry taxonomy.
- Separate consumer websites from infrastructure and service endpoints.
- Set a minimum sample per sector and company-size tier.

Pages

- Test homepage, article, product, documentation, and pricing templates.
- Record page role before outcome review.

Retrieval

- Compare raw HTTP, rendered browser, and named crawler user agents.
- Repeat requests across time and region to separate policy from transience.

Outcomes

- Pair delivery tests with real citation and source-selection observations.
- Report sector estimates with sample sizes and uncertainty intervals.

A proposed benchmark question

Which industries and page templates lose the most content between the browser experience and the response available to search, answer-time, and user-directed AI crawlers?

ABOUT THE RESEARCH

Method in brief

Sample

1,000 domains drawn from Tranco list 5674N, with 250 from each of four rank bands. Homepage only.

Collection

One request per domain on September 5, 2026. Research user agent. robots.txt checked before retrieval. JavaScript not executed.

Primary analysis

476 pages returning at least 50 body words. Spearman rank correlation with 10,000 seeded bootstrap resamples.

Observable signals first

Status, response size, body words, H1, JSON-LD, likely client rendering, redirects, and crawler-access rules.

Interpretation boundary

The study measures delivery and extractability, not whether an AI engine selected, ranked, quoted, or cited a page.

Open data

All 1,000 rows are retained, including failures. The reproducible analysis accompanies this report.

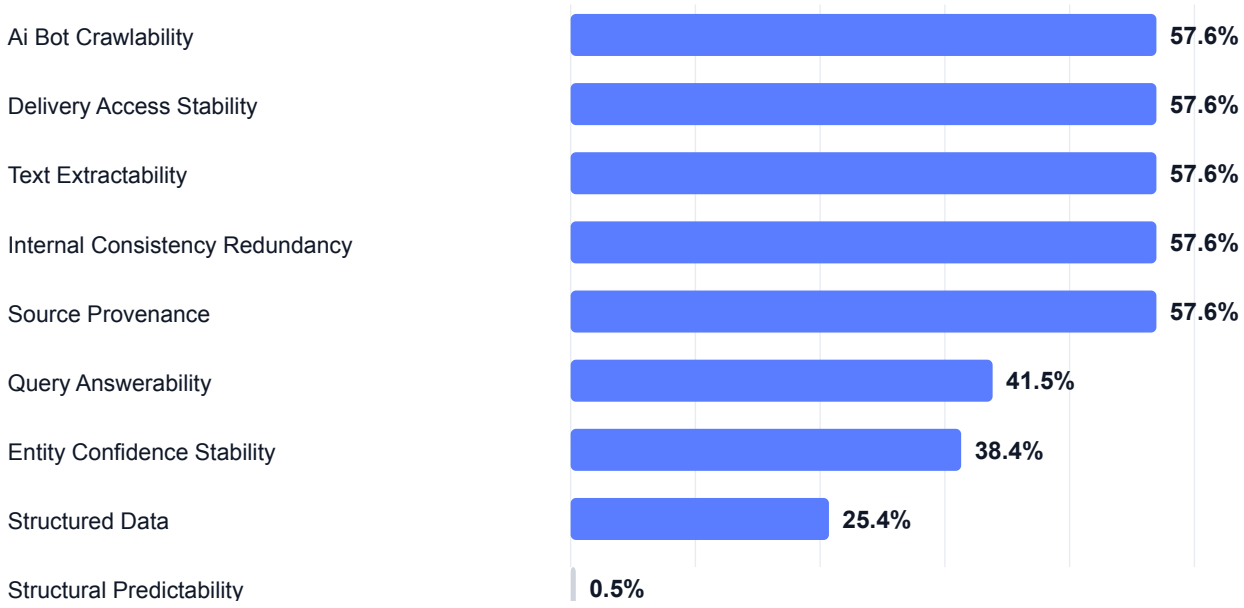

Exhibit 9. Coverage of recorded AI-score components across all rows. Component availability varies because not every check applied to every response. Cross-page context was not recorded and is omitted from the chart. Source: AIO Mapper study dataset.

CAVEATS AND ASSUMPTIONS

How to read the evidence

What the study supports

- A point-in-time estimate of what this research crawler received from the sampled homepages.
- Descriptive comparisons across rank bands and recorded page signals.
- A weak association between the two product-defined scores among qualifying pages.
- An operational taxonomy of the recorded delivery outcomes.

What remains outside scope

- Citation frequency, ranking, or source selection in AI products.
- Every crawler identity, geography, browser session, or retry.
- Industry-level estimates without a source taxonomy.
- Causal effects of schema, headings, robots changes, rendering, or page weight.

> **Interpretation principle:** Observable response signals carry the main claim. Composite scores provide supporting context and should not be treated as neutral facts.

Planned-versus-applied method notes

The applied analysis used HTML plus the word threshold because title presence was not stored. The named-crawler subgroup was recovered from the recorded crawlability component because the direct boolean field was blank. H2 is treated as less confirmatory because a live progress display showed an approximate unreadable rate when its threshold was set. Full details appear in the technical companion. [2]

RESEARCH RECORD

Sources and reproducibility

[1] AIO Mapper. 1,000-domain crawler readability dataset, September 2026. aiomapper.com/study-2026-09.csv

[2] AIO Mapper. How We Tested AI Crawler Readability Across 1,000 Domains. Technical companion

[3] AIO Mapper. The AI Visibility Debate Is Starting at the Wrong End. Findings article

[4] Tranco. List methodology. Tranco describes ranking inputs that include websites and infrastructural domains. tranco-list.eu/methodology

[5] OpenAI. Publishers and Developers FAQ. Guidance on OAI-SearchBot and inclusion in ChatGPT summaries and snippets. OpenAI Help Center

[6] Google Search Central. Optimizing your website for generative AI features on Google Search. Updated July 10, 2026. Google documentation

[7] Anthropic. Does Anthropic crawl data from the web, and how can site owners block the crawler? April 7, 2026. Claude Help Center

[8] Cloudflare. From Googlebot to GPTBot: Who's crawling your site in 2025. Includes analysis of AI-bot robots directives across top domains. Cloudflare Blog

[9] Rae et al. Scaling Language Models: Methods, Analysis and Insights from Training Gopher. The 50-word floor was adapted as a permissive document-length filter. arXiv:2112.11446

Reproducible files

The report package includes the original CSV, an enriched audit file, the computed metrics, and the analysis script used to produce every exhibit. Sector labels in the enriched file are exploratory and are not used in primary findings.

1,000 rows retained | 10,000 bootstrap resamples | Wilson intervals | Open data

AIO MAPPER

Before a page can rank, answer, or earn a citation, it has to reach the machine.

AI visibility diagnostics for the pages you publish.

aiomapper.com