Crawlito

Sample report · Illustrative site, measured from saved test pages

AI search readiness report: harborbakery.example

Scan date: 2026-11-11 · Pages checked: 5

Executive summary

We checked how AI search crawlers can reach and read harborbakery.example. The site scored 75 out of 100 on the published checks, with 100% coverage. The checks recorded 8 findings (1 high, 4 medium, 2 low, 1 informational). Start by reviewing the robots.txt rule that blocks Claude-SearchBot on /gift-cards. Each finding below lists why it matters, the exact change to make and how to confirm it.

Technical readiness

75/100

Check coverage: **100%** of applicable rule weight evaluated.

Category	Score
Access · 40% weight	77
Readable content · 30% weight	81
Structure · 20% weight	83
Discovery · 10% weight	33

What to fix first

Start with the most severe findings. Within each severity, address the finding with the widest observed reach first.

1. **robots.txt blocks AI search crawlers** · high · Server configuration · +9 readiness points if the whole check passes

   Open /robots.txt and find the group (User-agent line) that matches the listed bot, or the "*" group if it has none (Applebot uses the Googlebot group before the "*" group).

2. **Page is empty or holds only navigation and footer** · medium · Page content · +6 readiness points if the whole check passes

   Check whether the page is meant to have content; if it is a placeholder, add the real copy or remove it from navigation and the sitemap.

3. **Internal links point to pages that return errors** · medium · Page content · +3 readiness points if the whole check passes

   For each broken URL in the evidence, decide whether the page should exist.

Fix plan

7 scored checks failed on the checked pages. The table lists every one in a suggested order, starting from the current score of 75. Score after this step is the overall score from the published scorer once that step and every

step above it pass, with all other evidence unchanged.

Step	Check	Where the fix is made	Effort	Pages affected	Points	Score after this step
1	Search crawler policy	Server configuration	Quick change	Site level	+9	84
2	Content beyond navigation and footer	Page content	Larger project	1	+6	90
3	HTTP responses from sitemap entries	CMS and page templates	Template or config change	1	+3	93
4	Language alternate destinations	CMS and page templates	Template or config change	1	+2	94
5	Labels on crawlable links	CMS and page templates	Template or config change	1	+2	96
6	Responses from linked internal URLs	Page content	Larger project	Site level	+3	98
7	Duplicate content across URLs	Server configuration	Larger project	2	+2	100

The order is points per unit of effort. The effort label is a rough size, not a time estimate. Points are the score change if that check alone passed; each score is rounded once, so the points of single steps need not add up to the last figure. Site level marks a check that is not counted per page: a site-wide file or setting, or a check whose affected URLs are link targets. The plan describes the checked pages only and adds nothing to the score.

What the pages show

Each measure uses the pages, responses or images we could examine. The examples identify where to look next.

Successful HTTP responses

4 of 5 URLs

URLs returning HTTP 200–299 among those with a recorded server response.

Examples outside this count

- `https://harborbakery.example/old-menu`

Text available before JavaScript

3 of 3 pages

Pages whose server text is at least 50% of the browser text length. Includes paired views with at least 250 rendered characters. This compares length only.

Pages linked from the sample

1 of 3 pages

Distinct HTML destinations beyond the start page linked by another checked HTML page. Other pages on the site may also link to them.

Examples outside this count

- `https://harborbakery.example/cakes-copy`
- `https://harborbakery.example/gift-cards`

Search descriptions

1 of 4 pages

Successful HTML pages with a nonempty meta description. Search engines may also build snippets from the page text.

Examples outside this count

- `https://harborbakery.example/cakes`
- `https://harborbakery.example/cakes-copy`
- `https://harborbakery.example/gift-cards`

Typed structured data

3 of 4 pages

Successful HTML pages with at least one parseable JSON-LD block containing a named type. Use the structured data checks for field errors.

Examples outside this count

- `https://harborbakery.example/gift-cards`

Image alt attributes

1 of 2 images

Image elements with an alt attribute in the examined HTML. An empty attribute counts because it can correctly identify a decorative image. Review meaning in context.

Examples outside this count

- `https://harborbakery.example/`

Where issues overlap

Up to five observed URLs with multiple failed checks, ordered by the number of distinct checks.

URL	HTTP	Failed checks on this URL
`https://harborbakery.example/`	200	Labels on crawlable links · Language alternate destinations
`https://harborbakery.example/old-menu`	404	HTTP responses from sitemap entries · Responses from linked internal URLs

Which crawlers the site allows

The published rules at `https://harborbakery.example/robots.txt` were assessed for 4 distinct paths on the start origin. File response: HTTP 200. Server logs are needed to confirm visits. User-requested fetch behavior

varies by provider. Each agent is assessed with the robots.txt group that names it, otherwise with the `*` group. Applebot is the documented exception: without a group of its own it follows the Googlebot group, then `*`. The matched line is the first Disallow rule in robots.txt that blocked one of the assessed paths for that agent.

Agent or control	Purpose	Start page	Paths allowed	Matched robots.txt line	Blocked examples
OAI-SearchBot	Search	Allowed	4 of 4	—	—
Claude-SearchBot	Search	Allowed	3 of 4	`Disallow: /gift-cards`	`/gift-cards`
PerplexityBot	Search	Allowed	4 of 4	—	—
Googlebot	Search	Allowed	4 of 4	—	—
Bingbot	Search	Allowed	4 of 4	—	—
Applebot	Search	Allowed	4 of 4	—	—
ChatGPT-User	User requests	Allowed	4 of 4	—	—
Claude-User	User requests	Allowed	4 of 4	—	—
Perplexity-User	User requests	Allowed	4 of 4	—	—
GPTBot	Training controls	Blocked	0 of 4	`Disallow: /`	`/, /cakes, /cakes-copy`
ClaudeBot	Training controls	Allowed	4 of 4	—	—
Google-Extended	Training controls	Allowed	4 of 4	—	—
Applebot-Extended	Training controls	Allowed	4 of 4	—	—
CCBot	Training controls	Allowed	4 of 4	—	—
Meta-ExternalAgent	Training controls	Allowed	4 of 4	—	—
Amazonbot	Training controls	Allowed	4 of 4	—	—
Bytespider	Training controls	Allowed	4 of 4	—	—

The site's crawler instructions compared

We compared robots.txt, its Content-Signal lines, the sitemap, llms.txt and the canonical links and robots directives found on the 5 URLs checked in this audit with each other. A URL is reported as disallowed only when robots.txt disallows it for every search crawler we evaluate.

No disagreement between these instructions was observed in the checked pages.

Content signals in robots.txt

robots.txt declares no Content-Signal line.

Crawl delays in robots.txt

robots.txt declares no Crawl-delay line.

These are comparisons of what the site publishes, stated as observations. They do not show how a crawler acts on it, and each list is a sample of the URLs examined. This section is left out when no page could be read. This measurement does not change the score.

What a crawler that does not run JavaScript receives

We compared the main content area of the browser view with the server HTML on 3 pages whose browser view has at least 50 words of main content. Consent banners, dialogs and chat widgets are left out of both views. The median share of distinct main-content words in the browser view that are already present in the server HTML is **100%**.

On every compared page, all main-content words of the browser view and the compared page elements were already in the server HTML.

The share compares vocabulary, not layout or meaning, and describes the compared pages only. Where a word begins and ends depends on the script: text in scripts written without spaces, such as Chinese, Japanese or Thai, is split by dictionary, so its counts are approximate. It says nothing about crawlers that run JavaScript. This measurement does not change the score.

How much a machine must read

We measured the server HTML of 4 pages: its size, the share of it that is visible text, and how far into the source the main landmark (a main element or role="main") begins.

Median HTML size

1.5 kB

Largest page: 1.7 kB.

Median estimated tokens

about 206

Estimate for pages written mainly in Latin script: visible text characters divided by 4.

Median text share

53%

Share of the HTML bytes that is visible text.

Pages over 2 MB

0 of 4

Google documents that Googlebot reads the first 2 MB of an HTML file. Other crawlers publish no limit.

Pages with the lowest text share first.

URL	HTML size	Text share	Main content starts
`https://harborbakery.example/gift-cards`	444 bytes	24%	67% into the source
`https://harborbakery.example/`	1.7 kB	49%	42% into the source
`https://harborbakery.example/cakes`	1.5 kB	56%	39% into the source
`https://harborbakery.example/cakes-copy`	1.5 kB	56%	39% into the source

Sizes are of the decoded server HTML alone, without scripts, styles or images loaded separately. Only pages whose HTML was read are counted; responses too large to read are reported with the server answers. The token figure is an estimate, not a tokenizer count. This measurement does not change the score.

Do the page dates agree

We read the machine-readable dates of 4 checked pages: 0 declare a modification or publication date, and 0 answered with an ETag or Last-Modified header. 4 sitemap entries were examined; 0 of them declare a lastmod date.

No checked page declared modification dates more than 30 days apart.

The compared dates are the sitemap lastmod, the JSON-LD dateModified of the page's main entity and the meta modified time, and only on pages that declare at least two of them. The HTTP Last-Modified header is listed but not compared, because on many hosts it is the time of the last deploy and not of the last content change.

Dates are compared with other dates of the same page, never with the scan date, so this does not say whether the content is current. This measurement does not change the score.

Structured data compared with the visible page

We compared JSON-LD values with the visible text, title, description or image alt text of the same page on 3 pages with a parseable JSON-LD block. 3 values were checked.

Values not found as written in the visible text, title, description or image alt text of their page. Each distinct value is listed once, under the first page it was seen on; up to 8 are listed.

URL	Type	Field	Value in JSON-LD
`https://harborbakery.example/cakes`	Article	headline	Our story

Organization identity

The Organization markup on the checked pages declares one Organization name. Names: `Harbor Bakery`. URLs declared: `https://harborbakery.example/`. Only Organization and Corporation nodes at the top level of a block, in its @graph or given as a publisher are read; the URLs are listed and not compared.

A value can be absent for a sound reason, such as wording that differs or text shown only inside an image. Numbers are compared by their digits, so a different currency format or a rounded rating still counts as found. This compares text and does not judge the markup. This measurement does not change the score.

Findings and fixes

Access (2)

HIGH robots.txt blocks AI search crawlers

Policy files · 1 · high confidence · Rule `access.robots.search-bots-blocked`

Example pages

- `https://harborbakery.example/robots.txt`

Evidence

- bots: `Claude-SearchBot`
- rules[1].bot: `Claude-SearchBot`
- rules[1].matchedRule: `Disallow: /gift-cards`
- rules[1].blockedPaths: `/gift-cards`

- rules[1].blockedPathCount: `1`
- rules[1].homepageBlocked: `false`

Why it matters

OAI-SearchBot, Claude-SearchBot and PerplexityBot fetch content for their search services. Googlebot crawls for Google Search, including AI Overviews and AI Mode; Bingbot crawls for Bing and Copilot answers; Applebot crawls for Siri and Spotlight suggestions. A matching Disallow rule restricts access to the affected paths.

How to fix

1. Open /robots.txt and find the group (User-agent line) that matches the listed bot, or the "*" group if it has none (Applebot uses the Googlebot group before the "*" group).
2. Remove or narrow the Disallow line quoted in the evidence so public pages are allowed.
3. If you want search bots in but training bots out, add separate groups, e.g. "User-agent: OAI-SearchBot" with "Allow: /" and keep "User-agent: GPTBot" with "Disallow: /".
4. If you use Google-Extended to control training, manage its group separately. Allow Googlebot to fetch any pages you want it to read for Google Search.
5. Check that your CDN or hosting does not serve a different robots.txt than the one in your code.

How to verify

Test each affected path against the published robots.txt using the named search crawler. Confirm every path intended for search is allowed, then rerun the audit.

INFO robots.txt opts out of AI training crawlers (informational)

Policy files · 1 · high confidence · Rule `access.robots.training-bots-opted-out`

Example pages

- `https://harborbakery.example/robots.txt`

Evidence

- bots: `GPTBot`
- rules[1].bot: `GPTBot`
- rules[1].matchedRule: `Disallow: /`
- rules[1].blockedPaths: `/, /cakes, /cakes-copy`
- rules[1].blockedPathCount: `4`
- rules[1].homepageBlocked: `true`

Why it matters

A training or content-use control (GPTBot, ClaudeBot, Google-Extended, Applebot-Extended, CCBot, Meta-ExternalAgent, Amazonbot or Bytespider) is restricted on the sampled paths. This records a publishing preference. Search access is checked separately for the search crawlers.

How to fix

1. No change is needed if the opt-out is intentional.
2. Confirm the search crawlers (OAI-SearchBot, Claude-SearchBot, PerplexityBot, Googlebot, Bingbot, Applebot) have their own allowed groups so the training opt-out does not catch them through a shared "*" rule.
3. Google-Extended controls uses outside Google Search. Review its policy separately from Googlebot access.

How to verify

Test the affected paths for each named training control and each search crawler. Confirm both results match your intended policy.

Readable content (1)

MEDIUM Page is empty or holds only navigation and footer

Affected URLs · 1 · high confidence · Rule `content.page.empty`

Example pages

- `https://harborbakery.example/gift-cards`

Evidence

- contentChars: `10`
- totalChars: `106`
- excerpt: `Gift cards`

Why it matters

A page whose only text is the menu, header and footer gives a crawler nothing to describe or quote, and can dilute how the rest of the site is understood.

How to fix

1. Check whether the page is meant to have content; if it is a placeholder, add the real copy or remove it from navigation and the sitemap.
2. If content is loaded by scripts that failed, fix the data source or render it on the server.
3. Put the page's own content inside a <main> element so it is distinguishable from site chrome.

How to verify

Open the page with JavaScript disabled and confirm it shows its own heading and at least a paragraph of copy besides the menu and footer.

Structure (2)

MEDIUM Language alternates have incomplete or conflicting destinations

Affected URLs · 1 · high confidence · Rule `structure.hreflang.targets`

Example pages

- `https://harborbakery.example/`

Evidence

- problems: `Missing or relative destination for fr`

Why it matters

Search engines use language alternates to select a localized version of a page. A missing or conflicting destination makes the intended version ambiguous.

How to fix

1. Review the hreflang links in the page head. Give each language and region combination a single absolute URL, including the scheme and host.

How to verify

Inspect each flagged language in the page source and open its absolute destination to confirm the intended localized page loads.

LOW Links have no text or label

Affected URLs · 1 · high confidence · Rule `structure.links.unnamed`

Example pages

- `https://harborbakery.example/`

Evidence

- linksWithoutLabels: `1`
- destinations: `https://harborbakery.example/cakes`

Why it matters

A link label tells readers and crawlers what they will find at its destination. An empty link omits that context.

How to fix

1. Add descriptive text to the affected links. Give an image-only link alt text that describes its destination.

How to verify

Inspect the listed links in the page source and confirm each has a label that describes where it goes.

Discovery (3)

MEDIUM Internal links point to pages that return errors

Affected URLs · 1 · high confidence · Rule `discovery.links.broken-internal`

Example pages

- `https://harborbakery.example/old-menu`

Evidence

- status: `404`
- linkedFrom: `https://harborbakery.example/`
- sourceCount: `1`

Why it matters

Crawlers follow internal links to discover pages. Links that end in 4xx or 5xx errors waste the crawl and leave the linked content unreachable.

How to fix

1. For each broken URL in the evidence, decide whether the page should exist.
2. If it moved, add a permanent (301) redirect to the new address and update the links shown as sources.
3. If it was removed, delete or replace the links in the source pages and drop it from the sitemap.

How to verify

Open each listed URL and confirm it returns 200 or redirects to a page that does, and that the source pages no longer link to error pages.

MEDIUM The sitemap lists URLs that return errors

Affected URLs · 1 · high confidence · Rule `discovery.sitemap.error-target`

Example pages

- `https://harborbakery.example/old-menu`

Evidence

- httpStatus: `404`
- sitemapEntriesFetched: `4`

Why it matters

The sitemap directs crawlers to pages the site wants discovered. An error response at a listed URL prevents them from reading that page.

How to fix

1. Restore each listed page that should still exist. Remove obsolete addresses from the sitemap generator and publish the updated sitemap.

How to verify

Fetch the sitemap again and confirm every flagged address either returns its intended page or has been removed.

LOW Several URLs serve the same content without a shared canonical

Affected URLs · 2 · high confidence · Rule `discovery.pages.duplicate`

Example pages

- `https://harborbakery.example/cakes`
- `https://harborbakery.example/cakes-copy`

Evidence

- duplicateUrls: `https://harborbakery.example/cakes, https://harborbakery.example/cakes-copy`
- count: `2`
- otherSamples[1].duplicateUrls: `https://harborbakery.example/cakes, https://harborbakery.example/cakes-copy`
- otherSamples[1].count: `2`

Why it matters

When the same text is reachable at several addresses and none is marked as the main one, crawlers split their attention between copies and may pick the wrong address to show.

How to fix

1. Pick one URL for each group of duplicates listed in the evidence.
2. Redirect the others to it with a 301, or add <link rel="canonical"> on every copy pointing to the chosen URL.

3. Fix the source of the duplicates (tracking parameters, trailing-slash variants, print versions) so new copies are not created.

How to verify

Open each duplicate URL and confirm it either redirects to the chosen URL or declares it as canonical.

Developer appendix

The fixes above, grouped by where the change is usually made, so each part can go to the person who owns it.

Server configuration (3)

robots.txt, redirects, status codes and response headers on the web server or hosting platform.

robots.txt blocks AI search crawlers · high

Several URLs serve the same content without a shared canonical · low

robots.txt opts out of AI training crawlers (informational) · info

CMS and page templates (3)

Base layout, page templates, theme or framework settings that apply to many pages at once.

The sitemap lists URLs that return errors · medium

Language alternates have incomplete or conflicting destinations · medium

Links have no text or label · low

Page content (2)

Individual pages, links and files that an editor changes one by one.

Page is empty or holds only navigation and footer · medium

Internal links point to pages that return errors · medium

Check by check

Each rule appears once, even when several pages share a finding. A points figure shows the change in the overall score if that whole check passed with all other evidence held fixed. Rule weights are relative within each category. Unknown checks lack evidence. A check marked not applicable has no eligible subject in the sample.

Access · 40%

Check	Result	Rule weight
Start page HTTP response	Passed	10
Start page redirect destination	Passed	2
Search crawler policy	Failed · +9 pts	10
User-requested fetch policy	Passed	6

Check	Result	Rule weight
Training crawler policy	Advisory	Informational
Indexing and snippet directives	Passed	8
Firewall challenges	Passed	8
Optional llms.txt file	Passed	Informational

Readable content · 30%

Check	Result	Rule weight
Main text in server HTML	Passed	10
Raw and rendered text lengths	Passed	8
Document language	Passed	3
Content beyond navigation and footer	Failed · +6 pts	5

Structure · 20%

Check	Result	Rule weight
Document title	Passed	4
Main page heading	Passed	3
Heading order	Passed	2
Main content landmark	Passed	2
Canonical URL consistency	Passed	3
Labels on crawlable links	Failed · +2 pts	2
Language alternate destinations	Failed · +2 pts	2
JSON-LD syntax	Passed	3
Fields in supported structured data	Passed	3

Discovery · 10%

Check	Result	Rule weight
Sitemap discovery	Passed	4
HTTP responses from sitemap entries	Failed · +3 pts	3
Responses from linked internal URLs	Failed · +3 pts	3
Duplicate content across URLs	Failed · +2 pts	2

Methodology and limits

Each category earns the weight of its passed rules divided by the weight of all rules it could evaluate. The overall score combines category scores using access 40%, readable content 30%, structure 20% and discovery 10%. Categories without evaluated weight are left out and the remaining weights are scaled proportionally.

Check coverage is the percentage of applicable rule weight evaluated. Scoring requires at least 80% coverage and a successful start page response. Each scored rule counts once across the sample. The affected URLs and page measurements show the extent of a problem.

Training preferences, the optional llms.txt file and the descriptive page measurements carry no additional points.

Additional measurements. The fix plan orders failed checks by score points divided by an effort weight (quick change 1, template or config change 2, larger project 4); the weights only decide the order, and a step shows Site level when its check is not counted per page. The JavaScript comparison uses pages whose browser view has at least 50 words of main content (the main landmark, else a single article element, else the body without navigation, header, footer and sidebars), leaves out consent banners, dialogs and chat widgets, and reports the share of distinct main-content words of the browser view that are present in the server HTML; word boundaries depend on the script. Text share is visible text as a share of the HTML bytes. Estimated tokens are visible text characters divided by 4, a rule of thumb, not a tokenizer, given only for pages written mainly in Latin script. Page weight counts only pages whose HTML was read. Response times come from one request per page sent from our server. The crawler instructions section compares robots.txt, Content-Signal lines, the sitemap, llms.txt, canonical links and robots directives with each other, reports a URL as disallowed only when robots.txt disallows it for every search crawler evaluated, and is left out when no page could be read. Modification dates of one page are reported as disagreeing when they lie more than 30 days apart. The compared dates are the sitemap lastmod, the JSON-LD dateModified of the page's main entity and the meta modified time, and only on pages that declare at least two of them. The HTTP Last-Modified header is listed but not compared, because on many hosts it is the time of the last deploy and not of the last content change. Sitemap figures describe the sitemap entries examined, which can be fewer than the sitemap lists. Structured data values are compared with the visible text, title, description or image alt text of the same page, and Organization consistency is about the Organization name only. The 2 MB figure is the size Google documents for Googlebot; every other threshold here is our own heuristic. None of these measurements affects the score.

The score measures these published checks, not the probability of being cited by ChatGPT, Perplexity or Google AI.

Results reflect the site at the time of the scan. Pages behind logins, blocked by the site, or beyond the page limit are not covered. The page limit for this audit was 500 pages. Rule version 2026.11.3.

How to measure business outcomes

Use Search Console for impressions, clicks and average position by query and page. Use verified server logs for crawler visits, referral analytics for visits and conversions, and dated prompt observations for citations. These sources supply outcomes beyond this crawl.

Reference guidance · Google AI search eligibility · Indexing and snippet controls · Image alternatives