

GEO Casa · deck 1 of 4

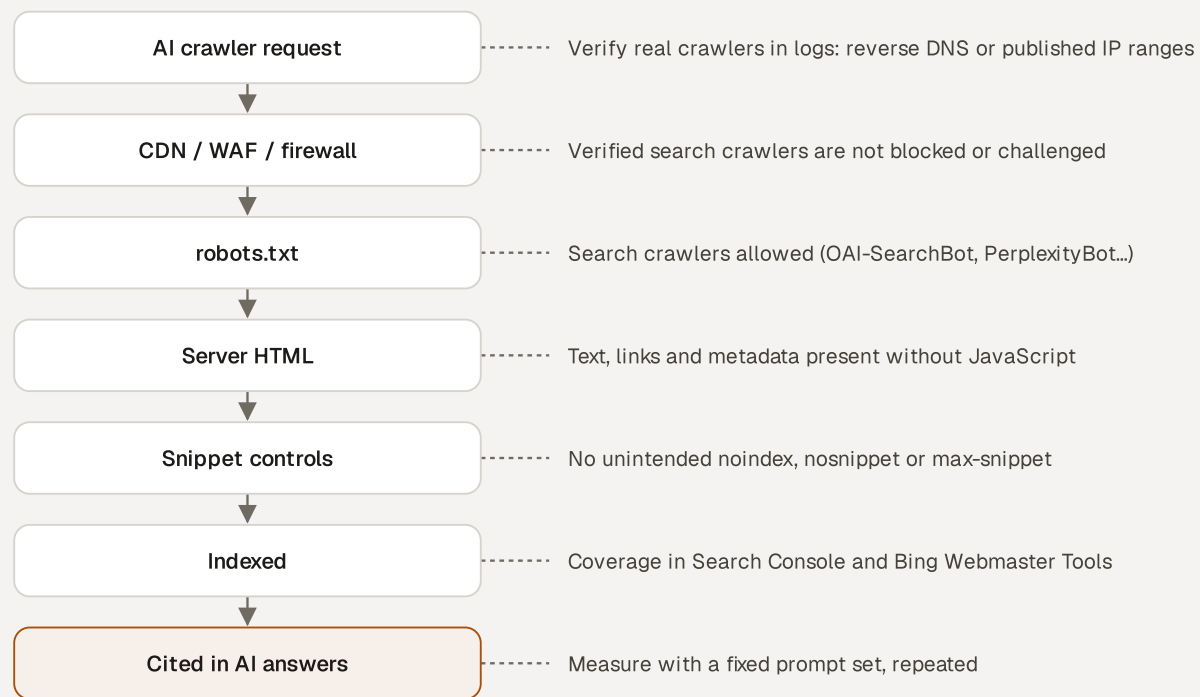
GEO by website

WordPress · JavaScript apps · online stores · docs · technical checks

GEO



How many gates before AI sees a page?



Work top to bottom. Content work on a page AI can't fetch is wasted.

[Read the full guide](#)



robots.txt: three common surprises

When rules conflict, including wildcards, Google uses the least restrictive rule. It does not support crawl-delay.

• Official · Google

A page blocked by robots.txt can still appear in results, without a description, if other pages link to it.

• Official · Google Search Central

In a 2026 audit of 130 company sites, the authors call blocking search and training crawlers together the most common and most damaging mistake.

• Observation · 2026-08 · DATA17 AI Research Lab · 解数智能

robots.txt controls crawling. To keep a page out of results, allow crawling and use noindex.



Try it: block a crawler, see what you lose

Search crawlers and training crawlers are separate decisions. Tap a token to block it.

OAII-SearchBot

Allowed

OpenAI · ChatGPT search

Not shown in ChatGPT search answers; may still appear as a navigational link.

● Official · OpenAI

GPTBot

Allowed

OpenAI · model training

Content not used for training; no effect on search.

● Official · OpenAI

PerplexityBot

Allowed

Perplexity · Perplexity search

Perplexity recommends allowing it to appear in results; the effect of blocking isn't spelled out.

● Official · Perplexity

Google-Extended

Allowed

Google · Gemini training and grounding

No effect on Google Search inclusion, including AI Overviews.

● Official · Google

CLaude-SearchBot

Allowed

Anthropic · Claude search indexing

May reduce visibility in Claude's answers.

● Official · Anthropic

CLaudeBot

Allowed

Anthropic · model training

Content not used for training.

● Official · Anthropic

Toggle demo. Each effect is quoted from the crawler owner's documentation.



Myth: Google-Extended and AI Overviews

? **Blocking Google-Extended keeps my pages out of Google's AI Overviews.**

False

Google-Extended covers Gemini training and grounding. AI Overviews are part of Search, so Googlebot rules and snippet controls such as nosnippet decide what they may use.

"Google-Extended does not impact a site's inclusion in Google Search nor is it used as a ranking signal in Google Search."

● Official · Google

[Read the full entry](#)



If robots.txt allows it, can the crawler get in?

Google lists "ensuring that crawling is allowed in robots.txt, and by any CDN or hosting infrastructure" among practices that still matter for AI features.

- Official · Google Search Central

From September 15, 2026, Cloudflare's defaults for new domains block training and agent bots on pages that show ads; search crawlers stay allowed.

- Official · Cloudflare

"Allowed in robots.txt" and "can access" are two different checks.



Try it: which crawler is genuine?

Anyone can type "GPTBot" into a user agent. Tap a line to see how to tell.

```
66.249.66.1 - - [28/Oct/2026:09:12:04] "GET /pricing/" 200 "Googlebot/2.1 (+http://www.google.com/bot.html)"
```

Genuine Reverse DNS gives a googlebot.com host, and the forward lookup returns the same IP.

```
203.0.113.45 - - [28/Oct/2026:09:13:51] "GET /pricing/" 200 "Mozilla/5.0; compatible; GPTBot/1.1"
```

Impostor The IP is not in OpenAI's published gptbot.json list. The name in the user agent proves nothing.

```
198.51.100.7 - - [28/Oct/2026:09:15:20] "GET /docs/" 200 "Mozilla/5.0; compatible; Claude-SearchBot/1.0"
```

Needs checking Match the IP against the list Anthropic links from its crawler help article before counting the visit.

```
192.0.2.10 - - [28/Oct/2026:09:16:02] "GET /" 403 "Mozilla/5.0; compatible; Baiduspider/2.0"
```

Needs checking Run a reverse DNS lookup. Baidu says genuine hosts end in .baidu.com or .baidu.jp. If genuine, the 403 is your firewall.

Simulated log lines with documentation IP addresses. Count only verified visits.

Google's crawlers can be verified with a reverse DNS lookup followed by a forward lookup, or against Google's published IP lists.

● Official · Google



What each status code tells Google

Crawlers act on the status code before they read a word of the page.

Status	What Google does	Basis
200, but an empty or error page	Not guaranteed to be indexed; reported as a soft 404	Official
301	Follows it; a strong signal to process the target	Official
302	Follows it; only a weak signal	Official
4xx (except 429)	Not indexed; indexed URLs are removed	Official
401 / 403 to slow crawling	No effect on crawl rate. Google says don't do it	Official
429 / 5xx	Crawling slows; indexed URLs are dropped if it lasts	Official

Source: Google's documentation on HTTP status codes.



Snippet controls reach AI Overviews

`nosnippet` prevents content from being used as a direct input for AI Overviews and AI Mode; `max-snippet` limits how much may be used.

- Official · Google Search Central

These settings "can be read and followed only if crawlers are allowed to access the pages that include these settings."

- Official · Google Search Central

A `nosnippet` left in a template quietly removes a page from AI answers. Search for it.



Sitemaps and IndexNow do different jobs

Tap a card.

 **XML sitemap**



A list of your URLs. At most 50,000 URLs or 50 MB per file; bigger sites use an index file.

 **<lastmod>**



Google uses it only when it is consistently accurate. It ignores priority and changefreq.

 **IndexNow**



Tells participating engines about added, updated or deleted URLs, up to 10,000 per request.

 **Google**



Not an IndexNow endpoint. Keep an honest sitemap for Google.

An IndexNow submission alerts participating engines, "but each engine makes its own decision about whether to index it."

• Official · [IndexNow.org](https://indexnow.org)



Myth: the Indexing API

? **Google's Indexing API gets any page indexed fast.**

False

It is limited to job postings and livestream video pages. Submissions go through spam detection, and abusing the API can get access revoked.

"The Indexing API can only be used to crawl pages with either JobPosting or BroadcastEvent embedded in a VideoObject."

• Official · Google Search Central

[Read the full entry.](#)



One preferred URL, every language linked

Redirects and rel="canonical" are strong canonical signals and sitemap inclusion is weak. Combining them raises the chance your preferred URL is chosen.

- Official · Google Search Central

If two pages don't both point to each other, hreflang tags are ignored. Each language version should list itself and all other versions.

- Official · Google Search Central

Make canonical tag, internal links and sitemap point at the same URL.



Walk-through: a page is not indexed

Open the URL in Search Console's URL Inspection tool, then answer in order.

1 Does the live test show crawling allowed?

⚠ Remove the robots.txt rule. If you want the page out of results, robots.txt is the wrong tool; use noindex.

2 Is indexing allowed (no noindex)?

⚠ Find noindex in the robots meta tag or X-Robots-Tag header. Templates, CMS settings and staging are the usual source.

3 Is Google's chosen canonical your URL?

⚠ If the pages are not duplicates, point canonical tag, internal links and sitemap at the URL you want.

4 Is the page clearly worth indexing?

⚠ Google doesn't guarantee indexing. Make the page original and useful, then request indexing once.

✓ **After each fix, run Test live URL, then request indexing for that one URL.**

[Read the full guide](#)



Myth: schema gets you cited

? **Adding schema.org markup makes AI search cite my pages more often.**

Unproven

No published result shows it. One 2026 experiment found no gain on pages already cited, and nobody has tested pages that aren't. Use markup for rich results and clear entity data, and keep it matching the visible page.

Google says no special schema.org structured data is needed to appear in AI Overviews or AI Mode, and markup should match the visible text.

● Official · Google Search Central

[Read the full entry.](#)

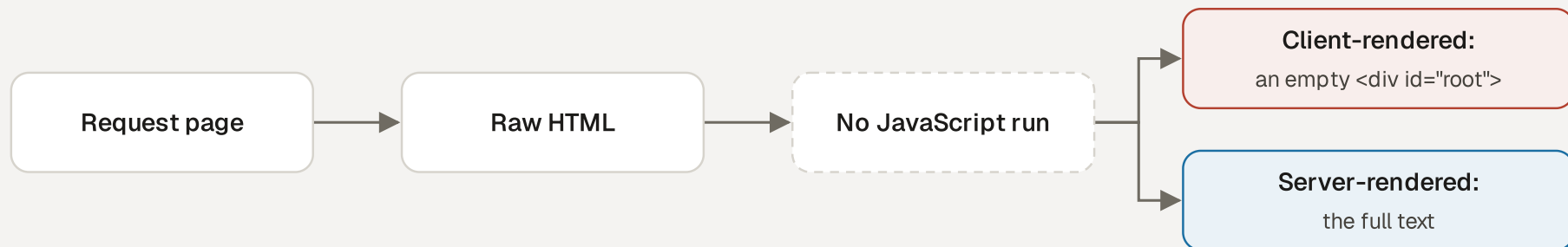


JavaScript sites: most AI crawlers read only the HTML

Browsers and Googlebot



Most AI crawlers (2024 test)



Vercel and MERJ (Dec 2024) found none of the major AI crawlers they tested ran JavaScript; Gemini did.

[Read the full guide](#)



JavaScript sites: what the docs say

Google still recommends server-side rendering or pre-rendering because "not all bots can run JavaScript."

• Official · Google Search Central

Google: "Dynamic rendering is a workaround and not a recommended solution." It recommends server-side rendering, static rendering or hydration.

• Official · Google Search Central

Vue's SSR guide warns that if an app starts with a spinner and fetches content via Ajax, "the crawler will not wait for you to finish".

• Official · Vue.js

A crawler that runs no JavaScript sees only the empty shell of the page.



Next.js, Vue, React teams

JavaScript sites: five checks

- 1 Fetch each template with JavaScript off
- 2 Render main content on the server: SSR, SSG or pre-rendering
- 3 Put links, metadata and structured data in that HTML
- 4 Serve bots the same page users see
- 5 Re-check after every release

Assume the crawler runs no JavaScript.

[Read the full guide](#)



WordPress: which setting hides the whole site?

"Discourage search engines from indexing this site" adds `noindex,nofollow` to every page that calls `wp_head()`.

- Official · WordPress.org

Google lists crawling allowed "by any CDN or hosting infrastructure" among practices that still matter for AI features.

- Official · Google Search Central

On WordPress, besides `robots.txt`, the Reading setting can add `noindex` to every page.



Site owners, agencies, hosts

WordPress: the route in six steps

- 1 Untick "Discourage search engines"
- 2 Read the robots.txt you actually serve; allow AI search crawlers
- 3 Check security plugin, host and CDN bot settings
- 4 Keep one XML sitemap and submit it
- 5 One plugin for structured data, matching the page
- 6 Keep the main text in the HTML, not in builder scripts

Most WordPress problems turn out to be settings.

[Read the full guide](#)



Online stores: feed, markup and page must agree

In Microsoft's Copilot example, a product makes the top three because its feed shows a competitive price and stock.

- Official · Microsoft Advertising

Google says using both page markup and a Merchant Center feed maximizes eligibility for product results.

- Official · Google Search Central

If feed, markup and page disagree on price, an AI has three answers to choose from.



Store owners and e-commerce teams

Online stores: six moves

- 1 Know the three sources: crawled pages, feeds, live site
- 2 Let search crawlers reach product pages
- 3 Mark up products to match what the page shows
- 4 Submit feeds; keep price and stock in sync
- 5 Answer buyer questions on the page
- 6 Earn real reviews; never fake them

Feed, markup and page should show the same price and stock.

[Read the full guide](#)



Docs sites: what the platforms already do

GitBook serves /llms.txt and /llms-full.txt, and appending .md to any published page URL returns that page as Markdown.

• Official · GitBook

Google Search ignores llms.txt: maintaining one "will neither harm nor help your site's visibility or rankings in Google Search".

• Official · Google Search Central

In Ahrefs' May 2026 logs of 137,210 domains, 97% of the llms.txt files published received no requests that month.

• Observation · 2026-05 · Ahrefs

Take llms.txt if your platform gives it free. Don't expect it to move rankings.



Docs sites and developer products

Docs get quoted one block at a time.

- ✓ Let search crawlers and user fetchers reach the docs
- ✓ A clean Markdown version of each page
- ✓ Code samples that run on their own
- ✓ Full text of every page in the server HTML
- ✓ One canonical URL per version; "latest" obvious
- ✓ A dated changelog; llms.txt only if cheap

If curl can't see your code sample, neither can most AI crawlers.

[Read the full guide](#)



AI agents use your site like a visitor

"Agents can view your website in 3 primary ways: screenshots, raw HTML, and the accessibility tree."
Modern agents combine them.

• Official · web.dev (Google)

"Everything we suggest to make a site 'agent-ready' also makes sites better for humans."

• Official · web.dev (Google)

Use real buttons and links, label form fields, and keep the layout stable.



The access audit: six checks per URL

Run it on a handful of key URLs, not the whole site.

- ☒ Indexed and eligible for a snippet
- ☒ Not blocked by CDN, firewall or hosting
- ☒ Main content in the server HTML
- ☒ Search crawlers allowed in robots.txt
- ☒ No unintended noindex, nosnippet or AI exclusion
- ☒ Verified crawler visits in the logs

Passing all six makes citations possible. It doesn't guarantee them.

[Read the full guide](#)



Minimum viable GEO: a few hours a week

For one person running a site alone.

- ✓ Week 1: confirm crawlers can reach and quote you
- ✓ Week 2: make your five key pages clear and complete
- ✓ List 20 questions customers ask
- ✓ Each week, improve one page that answers one
- ✓ Each month, ask two or three assistants and record
- ✓ Skip special AI files, mass rewrites, bought mentions

Spend the first hour on access. If the site is blocked, no amount of content helps.

[Read the full guide](#)



What to check before writing anything?

Check every access gate is open first. Each slide condenses a full guide. Read the guide for the evidence behind it.

[WordPress](#)[Online stores](#)[JavaScript sites](#)[Docs sites](#)[Technical leads](#)[Minimum viable GEO](#)[AI search access audit](#)[Verify AI crawlers](#)[robots.txt for AI crawlers](#)[Snippet controls](#)[Page not indexed](#)[Back to the overview deck](#)