

AI Crawler Access Checklist

AI companies send crawlers to read websites, and each does one of three jobs. Your robots.txt states which you let in, but a CDN, firewall or security plugin can overrule it without anyone deciding to. Most blocks are accidents, and they're quick to find.

From lesson 31 · The full course and the GEO audit: reyse.co.uk/geocourse · October 2026 edition · General education, not legal advice

Job	Crawlers	If you block it	Advice
Search: builds the index behind an AI search feature	OAI-SearchBot (ChatGPT), Claude-SearchBot, PerplexityBot	That tool can't use your pages in its search answers. OpenAI says sites that block OAI-SearchBot won't be shown in ChatGPT search answers.	Allow
User request: visits because a person asked	ChatGPT-User, Claude-User, Perplexity-User	OpenAI and Perplexity say robots.txt may not apply to these.	Allow
Training: collects pages that may train future models	GPTBot, ClaudeBot	Doesn't remove you from AI search answers.	Your choice
Google's AI Overviews and AI Mode	Googlebot	They're part of Google Search, so blocking Googlebot removes you from Search itself. Google-Extended is a separate setting for Gemini training and grounding that, Google says, doesn't affect Search.	Never block Googlebot
Copilot and Bing's AI answers	Bingbot	Copilot works from Bing's index, so blocking Bingbot removes you from both.	Never block Bingbot
Apple (Siri, Spotlight)	Applebot	With no Applebot rules of its own, Apple says Applebot follows your Googlebot rules.	Allow

Sources: OpenAI, Anthropic, Perplexity, Google, Microsoft and Apple's own crawler pages, checked October 2026.

EXAMPLE: BLOCK TRAINING ONLY, KEEP SEARCH OPEN

```
User-agent: GPTBot
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: *
Allow: /

Sitemap: https://www.your-agency.example/sitemap.xml
```

LESSON 3.1 · REYSE.CO.UK/GEOCOURSE/AI-CRAWLER-ACCESS

Let the Right AI Crawlers In: robots.txt, CDN and Firewalls

DO THIS FIRST Open yourdomain.co.uk/robots.txt and look for "Disallow: /" under User-agent: * or under OAI-SearchBot, Claude-SearchBot or PerplexityBot.

CHECK

Read your robots.txt · It's always public, so this takes a minute.

- ☐ Type yourdomain.co.uk/robots.txt into your browser. "Not found" means nothing is blocked by it.
- ☐ Look for User-agent lines naming OAI-SearchBot, ChatGPT-User, GPTBot, Claude-SearchBot, Claude-User, ClaudeBot, PerplexityBot, Applebot or Google-Extended, and read the line under each. Disallow: / means blocked.
- ☐ Look for "User-agent: *" followed by "Disallow: /". That blocks every polite crawler, search engines included.
- ☐ Check every version of your site: www, without www, and any separate property search subdomain each have their own file.

Check the layer above: your CDN and firewall · These can refuse a crawler before it ever reads robots.txt.

- ☐ If your site uses Cloudflare, open Security and look for the AI crawler settings (Search, Agent and Training). Check Search is allowed.
- ☐ Check any security plugin (Wordfence, Sucuri and others) for rules that block bots or "unknown automated visitors".
- ☐ Check for UK-only geoblocking or harsh rate limits. AI crawlers may not visit from UK addresses.

Check who really visits · The only way to know the guard is letting crawlers through.

- ☐ Send your web provider the crawler log request from lesson 2.1.
- ☐ A 200 response means the page was served. 403 means refused. 429 means too many requests.
- ☐ Anyone can fake a crawler's name, so a technical person can check visits against the IP lists OpenAI, Anthropic and Perplexity publish.

Ask the assistants to open a page

- ☐ Paste one of your page addresses into ChatGPT, Claude and Perplexity.
- ☐ Ask: "Open this page and tell me what fees it lists."
- ☐ A refusal or error can point to a block, though it can have other causes. Repeat on two or three pages.

Ask AI to read the file

AI PROMPT Here is my robots.txt. Which AI crawlers does it allow and block? Explain in plain English, separate search crawlers, user-request crawlers and training crawlers, and flag anything that looks accidental, such as a blanket block.

FIX

Write a three-line policy first · Decide on purpose, write it down, and nobody can undo it by accident.

- ☐ Search crawlers: allow.
- ☐ User-request crawlers: allow (robots.txt may not stop them anyway).
- ☐ Training crawlers: your business choice. Some agencies block them on principle; others allow them, because an assistant answering from memory can only name a business it learned about. Neither choice has been shown to change mentions for agents.
- ☐ Save it with the date and who approved it. Review twice a year.

Fix robots.txt, carefully · One wrong line can hide your whole site, so keep a copy of the old file first.

- ☐ Remove any blanket "Disallow: /" you didn't intend.
- ☐ To block training only, add "User-agent: GPTBot" then "Disallow: /", a blank line, then the same for ClaudeBot.
- ☐ Don't add Google-Extended to that block unless you've decided to: Google says it also covers grounding in Gemini.
- ☐ Never block Googlebot or Bingbot as an "AI block". That removes you from ordinary search too.
- ☐ WordPress: Yoast has a file editor under Tools; Rank Math has Edit robots.txt under General Settings.

Fix the CDN or firewall

- ☐ In Cloudflare, keep Search crawlers allowed and make a deliberate choice about Agent and Training. Be careful: blocking Training can also catch crawlers that do both jobs, such as Googlebot and Bingbot.
- ☐ In a security plugin, allow the search crawlers you chose by name, and by IP list where the tool supports it.
- ☐ Keep your attack protection on. The aim is to let the right visitors through, not to open the doors.

- Ask your web provider in writing** · On many managed property websites you can't edit these settings yourself.
- ☐ Ask what robots.txt says, who can change it, and how long a change takes.
 - ☐ Ask whether any firewall, CDN or bot protection blocks OAI-SearchBot, Claude-SearchBot, PerplexityBot or Bingbot.
 - ☐ Ask for a log of refused requests from those crawlers.

Let AI draft it

AI PROMPT Write a robots.txt that allows OAI-SearchBot, ChatGPT-User, Claude-SearchBot, Claude-User, PerplexityBot, Perplexity-User, Applebot, Bingbot and Googlebot, and blocks GPTBot and ClaudeBot. Keep this sitemap line: [paste]. Explain each line.

Then: After any change, reload robots.txt in a browser, re-run the open-a-page test, and ask your provider for a fresh log after a week. OpenAI says changes take about 24 hours; other systems vary.

robots.txt is a request, not a lock

Reputable crawlers follow it. User-request fetchers may not. It's also a public file, so never list anything private in it.

It's not a legal shield either. The UK government's March 2026 report on copyright and AI made no change to the law, so there's no legislated opt-out from AI training. Treat blocking training crawlers as a business choice.

If you have no rules for Applebot but do have rules for Googlebot, Apple says Applebot follows the Googlebot rules, so a rule written for Google can affect Siri and Spotlight.

Your policy and checks

Item	What you found or decided	Date
Search crawlers		
User-request crawlers		
Training crawlers		
robots.txt (www and non-www)		
CDN AI settings (Search, Agent, Training)		
Security plugin or firewall		
Crawler log from your provider		
Approved by		

Rather have us do this for you? Our managed GEO service takes it off your hands: reyse.co.uk/products/geo