interweb

Generated on 15/11/2026 00:12

KNOWLEDGE BASE · PUBLIC DOCUMENT

AI bots crawling your site: robots.txt and how to block them

AI for Your Website

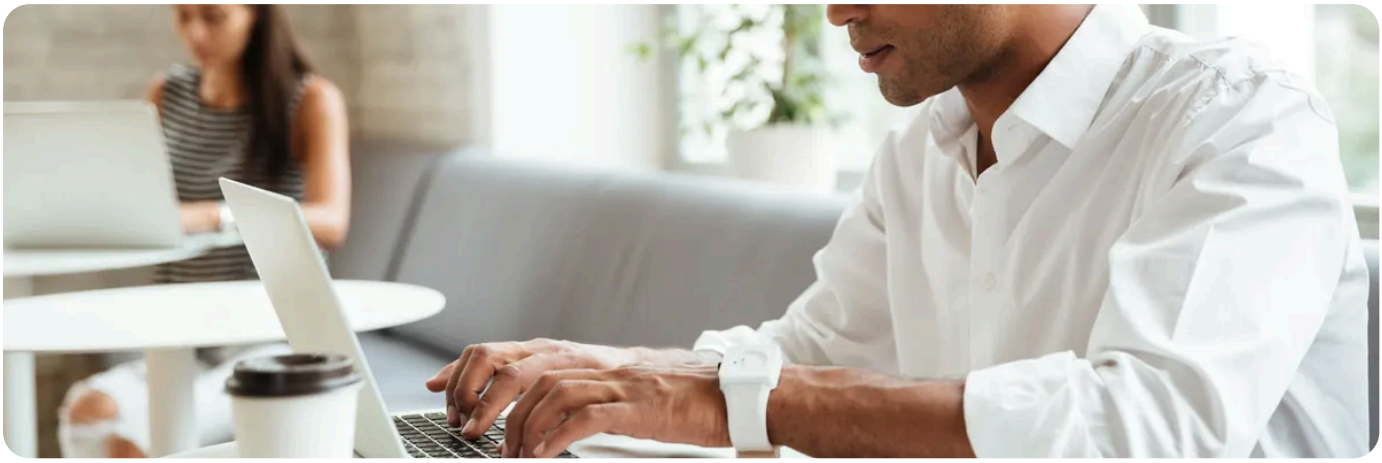
CODE	VERSION	COUNTRY · LANGUAGE
RLX-AO-KB-950	v2026.10.10	Angola · English

Interweb Angola
RLX Angola, Lda. · Group RLX Holdings
Tax ID (NIF) 500 260 5724 · Commercial registration no. 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/950

AI bots crawling your site: robots.txt and how to block them

AI for Your Website · Created on **October 10, 2026** · 2 min read



AI providers use robots (*crawlers*) that walk through websites collecting text, either to train models or to answer questions in real time. If you do not want that, you can ask them to stay out with **robots.txt**, and you can **refuse access on the server** by the name they identify themselves with. The first is a polite request; the second is a closed door, but it only works against those who tell the truth about who they are.

First, see who is there

1. **Open the site's access logs** in cPanel and look for the "user agent" field (the name the visitor introduces itself with).

2. **Look for names you recognise.** Providers publish the names of their robots in their documentation: for example GPTBot (OpenAI), ClaudeBot (Anthropic), CCBot (Common Crawl) and PerplexityBot. Names change, so always confirm in each one's documentation.

3. **Decide what you want.** Blocking means you may stop appearing in those services' answers. Not blocking means accepting that your text is read.

Asking with robots.txt

The `robots.txt` file sits at the root of the domain (`asuaempresa.ao/robots.txt`). For each robot, one line saying who it is and one saying what it may not see:

```
User-agent: GPTBot
Disallow: /

User-agent: ClaudeBot
Disallow: /
```

```
User-agent: CCBot
Disallow: /
```

Google has a separate name, `Google-Extended`, which, according to Google's public guidelines, controls whether content is used for its generative AI products. **Do not block** `Googlebot`, which is what indexes the site in search. See **what robots.txt and the sitemap do**.

Refusing on the server

robots.txt is voluntary: each robot complies if it wants to. To actually refuse, on a shared hosting site that uses the `.htaccess` file, you can deny access by name:

```
RewriteEngine On
RewriteCond %{HTTP_USER_AGENT} (GPTBot|ClaudeBot|CCBot) [NC]
RewriteRule .* - [F,L]
```

Back up .htaccess before editing it,
and check that the site still opens. A syntax error makes the whole site return a 500 error. If the site already has other rules, add these without deleting the existing ones. See **website errors explained**.

Neither robots.txt nor blocking by name stops those who lie.
Some robots present themselves as an ordinary browser. For those, the road is blocking by IP address or putting a protection layer in front of the site, such as Cloudflare. See **blocking an IP address** and **what Cloudflare is**.

If the issue is load, not privacy, first measure how much they request. A well-behaved robot rarely weighs on the server; an aggressive one can, and that one does not obey robots.txt.

NEED A HAND?

Is your site slow or throwing errors because of odd traffic? Open a ticket with the time and the address and we will see what the logs show.

Open a support ticket →

SEE ALSO

robots.txt and sitemap.xml: what they do and how to check them →

Blocking an IP address from visiting your site

Cloudflare: what it is, what changes, and what stays the same

interweb

Interweb Angola

RLX Angola, Lda. · Group RLX Holdings

TAX ID (NIF)	500 260 5724
COMMERCIAL REGISTRATION NO.	26464-25/250724
REGISTERED OFFICE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACT	info@interweb.ao · +244 942 011 191 · www.interweb.ao
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-AO-KB-950 · v2026.10.10 · Generated on 15/11/2026 00:12