ARKCES | CHEAT SHEET

robots.txt for AI crawlers

Which bot does what, and copy-ready rules. Edition: November 2026.

Rule one: blocking a search bot removes you from that tool's answers. Blocking a training bot only opts you out of model training. Decide them separately.

User agent	Company	Purpose	Block it and...
`Googlebot`	Google	Search, including AI Overviews and AI Mode	You leave Google Search
`Google-Extended`	Google	Token: Gemini training and grounding in Gemini Apps and Vertex AI	No effect on Search inclusion or ranking (per Google)
`Bingbot`	Microsoft	Bing and Copilot	You leave Bing and Copilot answers
`OAI-SearchBot`	OpenAI	ChatGPT search	You leave ChatGPT search answers
`GPTBot`	OpenAI	Model training	Training opt-out only
`ChatGPT-User`	OpenAI	Fetches a page a user asked for	OpenAI says robots.txt may not apply
`Claude-SearchBot`	Anthropic	Search results in Claude	You leave Claude search results
`ClaudeBot`	Anthropic	Model training	Training opt-out only
`Claude-User`	Anthropic	Fetches a page a user's question needs	Claude can't fetch your page for users
`PerplexityBot`	Perplexity	Perplexity search index (not training, per Perplexity)	You leave Perplexity answers
`Perplexity-User`	Perplexity	User-requested fetches	Perplexity says it generally ignores robots.txt
`Applebot-Extended`	Apple	Token: Apple model training	No effect on Apple search features

Sources: OpenAI bots documentation, Anthropic crawler help article, Perplexity crawler documentation, Google common crawlers documentation, Apple Applebot documentation. Changes to robots.txt can take around 24 hours to be picked up (OpenAI, Perplexity).

Copy-ready policies

A. Allow everything (most businesses)

```
User-agent: *
Allow: /

Sitemap:
https://example.co.za/sitemap.xml
```

B. Search yes, training no

```
User-agent: GPTBot
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: Google-Extended
Disallow: /

User-agent: Applebot-Extended
Disallow: /

User-agent: *
Allow: /

Sitemap:
https://example.co.za/sitemap.xml
```

OAI-SearchBot, Claude-SearchBot, PerplexityBot, Googlebot and Bingbot fall under the last rule and stay allowed.

C. Keep private areas out (any policy)

```
User-agent: *
Disallow: /wp-admin/
Allow: /wp-admin/admin-ajax.php
Disallow: /?s=
Disallow: /cart/
Disallow: /checkout/
```

robots.txt is public and is not security. Protect private pages with a login.

Check it works

- [ ] robots.txt loads at /robots.txt with status 200
- [ ] Search Console robots.txt report shows no errors
- [ ] CDN or firewall "AI bot" settings match your policy
- [ ] Server logs show search bots getting 200 responses
- [ ] Verify bot IPs against each company's published ranges before trusting a user agent

More: the full robots.txt guide, the AI Search Readiness Playbook and AI search optimisation services at arkces.co.za.