Customize your robots.txt

Last modified on 10/01/2026 3:47 pm EDT

The robots.txt file lets you tell AI crawlers and search indexing crawlers whether they're allowed to crawl your public knowledge base. You can customize the robots.txt to block specific crawlers from crawling your knowledge base. You can ignore this file if your knowledge base is completely private, since the login process prevents tools from crawling your site.

Most common use
Customize the robots.txt file if you want to prevent AI companies from training their models on your knowledge base content.

KnowledgeOwl generates a robots.txt file for all knowledge bases automatically. Here's what our default robots.txt file looks like:

```
User-agent: *
Disallow: /api/
```

If you have set your **SEO settings** to **Generate a publicly available sitemap** checked, robots.txt also includes that, for example:

```
User-agent: *
Disallow: /api/

Sitemap: https://myknowledgebase.knowledgeowl.com/help/sitemap.xml
```

How to customize robots.txt

To customize your robots.txt file:

1. Go to **KB settings > Domain.**
2. If this is your first time customizing your robots.txt file:
   a. Select the **robots.txt** link in **Custom robots.txt** to open your robots.txt file in a separate tab:



   b. Copy the text from that separate tab.
   c. Paste it into the **Custom robots.txt** text box in your **Domain** settings.

Copyright © 2015 – 2025 Silly Moose, LLC. All rights reserved.
Powered by KnowledgeOwl knowledge base software.

3. Enter your new customizations in the **Custom robots.txt** text box.
4. Be sure to **Save** your changes.
5. Select the **robots.txt** link in **Custom robots.txt** in your **Domain** settings to verify that your robots.txt file looks as you'd like.

Check out the Quickstart below for more information on formatting your rules and to copy some rules to prevent the scraping used to train AI models.

Quickstart to robots.txt customizations

The most common customization our authors like to do is to block certain AI-related crawlers from accessing their knowledge base or to allow specific crawlers or bots to have access.

Allow examples

To allow a specific crawler access to your knowledge base, you specify the `User-agent` you want to have access and then set what you want to `Allow` (or allow them to scan).

To allow a User-agent to scan your entire knowledge base:

```
User-agent: {User-agent name}
Disallow: /api/
Allow: /
```

A crawler that has its own `User-agent` entry follows only that entry and ignores the `User-agent: *` rules, so keep the `Disallow: /api/` line from the default file in it. Crawlers you don't list follow the default rules, which already allow your whole knowledge base.

Disallow examples

To block a crawler from accessing your knowledge base, you specify the `User-agent` you want to restrict and then set what you want to `Disallow` (or prevent them from scanning).

To prevent a User-agent from scanning your entire knowledge base:

```
User-agent: {User-agent name}
Disallow: /
```

There's no single rule that blocks one category, because KnowledgeOwl article URLs don't include the category they're in.

Combining allow and disallow

Copyright © 2015 – 2025 Silly Moose, LLC. All rights reserved.
Powered by KnowledgeOwl knowledge base software.

For example, OpenAI uses separate User-agents for training and for ChatGPT:

- `GPTBot` **: Collects data that may be used to train OpenAI's AI models**
- `OAI-SearchBot` **: Finds pages to show and link to in ChatGPT's search results**
- `ChatGPT-User` **: Visits a page when someone using ChatGPT asks it to. OpenAI says robots.txt rules may not apply to it.**

If we wanted to keep our knowledge base out of OpenAI's training data but still have it show up and be cited in ChatGPT's answers, we'd disallow GPTBot and allow OAI-SearchBot. We'd add these entries to our Custom robots.txt:

```
# Block OpenAI's training bot
User-agent: GPTBot
Disallow: /

# Allow ChatGPT search, which shows and links to pages in ChatGPT answers
User-agent: OAI-SearchBot
Disallow: /api/
Allow: /
```

Block all AI training scraping

If your goal is to block the scrapers that are used to train AI models, here are the key entries you'll want to add as of October 2026:

Copyright © 2015 – 2025 Silly Moose, LLC. All rights reserved.
Powered by KnowledgeOwl knowledge base software.

```
# Block AI training crawlers (as of October 2026)
# Search and answer bots aren't listed here, so they stay allowed.

# OpenAI
User-agent: GPTBot
Disallow: /

# Anthropic
User-agent: ClaudeBot
Disallow: /

# Google (Gemini training and Gemini app grounding; doesn't affect Google Search)
User-agent: Google-Extended
Disallow: /

# Apple (doesn't affect Apple search results)
User-agent: Applebot-Extended
Disallow: /

# Common Crawl (open web archive widely used to train AI models)
User-agent: CCBot
Disallow: /

# ByteDance
User-agent: Bytespider
Disallow: /

# Meta (Meta AI search uses Meta-WebIndexer, which stays allowed)
User-agent: Meta-ExternalAgent
Disallow: /
```

AI companies run separate bots for training their models and for answering questions in their chat and search tools, and this list blocks only the training bots, so your knowledge base can still appear and be cited in answers from ChatGPT, Claude, Perplexity, and Google Search.

Learn more

For more information on the User-agent strings for common AI tools, check out these resources:

- **Chris Lever's User Agent String Database & Bot Directory: Includes quick filters for AI Crawlers versus search and other crawlers with detailed entries explaining whether the crawler respects robots.txt directives and descriptions of what each crawler does**
- **SearchEngineJournal's Complete Crawler List for AI User-Agents [December 2025]**
- **LLMS Central's Complete Guide to AI Bot User-Agents, which has some quick copy-paste examples for robots.txt**

Copyright © 2015 – 2025 Silly Moose, LLC. All rights reserved.
Powered by KnowledgeOwl knowledge base software.