

Hieronder vind je een overzicht van de instructies die je kunt toevoegen aan robots.txt om het gebruik van jouw content voor het trainen van AI-modellen te blokkeren.

Common Crawl: deze organisatie verzamelt trainingsdata voor AI-modellen die vervolgens voor iedereen gratis te gebruiken is als open data. Ook grote bedrijven als OpenAI gebruiken de data van Common Crawl om AI-modellen te trainen.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: CCBot

Disallow: /

OpenAI: het bedrijf achter ChatGPT en het bijbehorende taalmodel GPT, maar ook veel andere AI-modellen.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: GPTBot

Disallow: /

ChatGPT-toegang: als een ChatGPT-gebruiker vraagt om informatie op te zoeken op een website, proberen de servers van OpenAI een webpagina te bezoeken en in te laden. Deze wordt niet gebruikt als trainingsdata, maar enkel eenmalig om de vraag van die gebruiker in kwestie te beantwoorden.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: ChatGPT-User

Disallow: /

Google: naast het indexeren van websites voor zijn zoekmachine, heeft Google ook een bot die trainingsdata verzamelt voor zijn AI-modellen. Als je die blokkeert heeft dat geen effect op je vindbaarheid op Google.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: Google-Extended

Disallow: /

Facebook: Ook Facebook heeft zijn eigen taalmodel ontwikkeld en gebruikt hiervoor trainingsdata die het op internet verzamelt.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: FacebookBot

Disallow: /

Anthropic: dit bedrijf is bij het grote publiek nog niet heel bekend, maar is een belangrijke concurrent van OpenAI, die zich profileert als 'veiliger' dan andere AI-bedrijven. Het taalmodel van Anthropic heet Claude.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: anthropic-ai

Disallow: /

Perplexity: dit bedrijf heeft een AI-zoekmachine gebouwd, die zoekvragen probeert te beantwoorden door direct het antwoord te geven in plaats van een lijst linkjes. De benodigde kennis haalt een bot van websites op het internet.

Blokkeren? Voeg deze instructies toe aan robots.txt:

User-agent: PerplexityBot
Disallow: /

Wil je al deze bedrijven en hun bots blokkeren? Dan kun je onderstaande instructies kopiëren en in robots.txt zetten. Heb je nog geen robots.txt, dan kun je deze alsnog maken en uploaden naar de hoofdmap van je website, waardoor deze te vinden is op www.domeinnaam.nl/robots.txt.

User-agent: CCBot
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: Google-Extended
Disallow: /

User-agent: FacebookBot
Disallow: /

User-agent: anthropic-ai
Disallow: /

User-agent: PerplexityBot
Disallow: /