

AI-crawlers in Nederland: trainingsbot dicht, zoekbot open

Op de 500 populairste .nl-domeinen worden de trainingsbots van AI-bedrijven duidelijk vaker geblokkeerd dan de zoekindexbots die bepalen of een site in de antwoorden van ChatGPT, Perplexity en Google verschijnt. Een meting van robots.txt, peildatum 17 juni 2026.

Auteur: Matt Timmermans, Timmermans Media. Steekproef: de 500 populairste .nl-domeinen volgens de Tranco-ranglijst, met 426 leesbare robots.txt als noemer.

Inleiding

AI-assistenten als ChatGPT, Perplexity en Google halen hun antwoorden niet uit één bron. Sommige crawlers verzamelen tekst om taalmodellen te trainen, andere bouwen de zoekindex waaruit een assistent live citeert, en weer andere halen een pagina pas op het moment van de vraag op. Welke crawler je toelaat of blokkeert bepaalt dus telkens iets anders. Uit onze meting van de 500 populairste .nl-domeinen blijkt dat de blokkades sterk leunen op de trainingsbots, terwijl de zoekindexbots veel vaker open blijven staan. Een site die alleen een trainingsbot blokkeert, blijft volgens de werking van die crawlers zichtbaar in de zoekgebaseerde antwoorden van AI.

Kerncijfers

- 20,4% blokkeert minstens één AI-crawler.
- 6,6% blokkeert alleen een trainingsbot en laat de zoekindexbots open, en blijft daarmee zichtbaar in AI-zoek.
- 13,1% blokkeert minstens één AI-zoekindexbot en is in de zoekgebaseerde antwoorden van die assistent niet citeerbaar.
- 13,4% blokkeert Google-Extended maar niet Googlebot.
- 0,7% blokkeert Googlebot, de enige crawler die je via robots.txt uit Google en AI Overviews haalt.
- 5,9% heeft een llms.txt-bestand.

Welke bot je blokkeert, bepaalt iets anders

Niet elke AI-crawler doet hetzelfde. Een trainingsbot als GPTBot of ClaudeBot verzamelt tekst om een model te trainen. Blokkeer je die, dan kom je niet in de trainingsdata, maar dat zegt niets over de antwoorden die je vandaag in ChatGPT ziet. Want die antwoorden komen van een andere bot: de zoekindexbot, zoals OAI-SearchBot voor ChatGPT, PerplexityBot voor Perplexity en Claude-SearchBot voor Claude. Die bouwen de index waaruit de assistent live citeert. En dan is er nog de live-fetch-bot, zoals ChatGPT-User, die een pagina pas ophaalt op het moment dat een gebruiker erom vraagt.

In de meting leunen de blokkades sterk op de eerste soort. GPTBot is met 17,6% de meest geblokkeerde crawler, gevolgd door de andere trainingsbots. De zoekindexbots van ChatGPT en Claude worden duidelijk minder geblokkeerd: OAI-SearchBot 9,2% en Claude-SearchBot 8,5%. Het scherpst zie je het binnen één bedrijf, omdat dat de site constant houdt. GPTBot wordt door 75 sites geblokkeerd en OAI-SearchBot maar door 39, dus minstens 36 van de GPTBot-blokkeerders laten de zoekbot van ChatGPT openstaan. Bij Anthropic is het beeld hetzelfde: ClaudeBot wordt door 14,8% geblokkeerd, de bijbehorende Claude-SearchBot maar door 8,5%. Die sites blijven volgens de werking van de crawlers citeerbaar in de zoekgebaseerde antwoorden, ongeacht waarom ze de trainingsbot blokkeren.

Achter dat ene cijfer van 20,4% zitten dus twee verschillende configuraties. Een deel, 6,6%, blokkeert alleen een trainingsbot en blijft zichtbaar in AI-zoek. Voor wie daarmee uit AI wilde verdwijnen is dat schijnveiligheid, en voor wie alleen modeltraining wilde tegenhouden is het juist precies de bedoeling. Een groter deel, 13,1%, blokkeert wel een zoekindexbot en is daardoor niet citeerbaar in de antwoorden van die assistent. Het cijfer beschrijft de configuratie, niet de bedoeling erachter.

Welke organisaties de trainingsbot blokkeren en de zoekbot openlaten

Het patroon is geen abstractie. NOS en RTL blokkeren GPTBot en ChatGPT-User, maar laten OAI-SearchBot openstaan, en blijven daardoor citeerbaar in de zoekgebaseerde antwoorden van ChatGPT. Hetzelfde zie je bij Naturalis, NPO Radio 1, BNR, Ticketmaster, Tripadvisor en eBay: de trainingsbot gaat dicht, maar de zoekbot van ChatGPT blijft open. Wat de reden ook is, het effect is hetzelfde: ze houden de trainingsbot buiten de deur en blijven tegelijk vindbaar in ChatGPT-zoek. Voor wie zijn content niet aan modeltraining wil afstaan maar wel geciteerd wil worden, is dat een verdedigbare keuze. Wie met het blokkeren van GPTBot dacht ChatGPT helemaal te hebben buitengesloten, bereikt dat niet. Welke van de twee het is, valt uit een robots.txt niet af te lezen.

Google-Extended blokkeren haalt je niet uit AI Overviews

Ruim een op de acht sites, 13,4%, blokkeert Google-Extended maar niet Googlebot. Dat heeft niet het effect dat je zou verwachten. Google-Extended regelt of je content gebruikt wordt om Gemini te trainen en te gronden, maar het haalt je volgens de eigen documentatie van Google niet uit AI Overviews. Die putten namelijk uit de gewone zoekindex van Googlebot, en Googlebot wordt door slechts 0,7% geblokkeerd. Het gevolg is dat de bedrijven die Google-Extended blokkeren bijna allemaal nog gewoon in AI Overviews verschijnen. Wie daar wel uit wil, kan dat niet via Google-Extended doen. Google kondigde op 3 juni 2026 wel een aparte opt-out aan in Search Console voor AI Overviews en AI Mode, los van de gewone zoekresultaten en zonder dat het de ranking raakt. Die opt-out wordt voorlopig alleen getest onder een beperkte groep site-eigenaren in het Verenigd Koninkrijk, met een bredere uitrol aangekondigd, dus voor Nederlandse sites is hij nog niet beschikbaar.

Googlebot blijft onaangeroerd

Tegenover de 17,6% voor GPTBot staat 0,7% voor Googlebot. Vrijwel niemand blokkeert de crawler die de gewone Google-zoekresultaten en, via diezelfde index, AI Overviews voedt. AI-training is het strijdtoneel geworden, terwijl de zoekindex van Google vrijwel onaangeroerd blijft. Dat verklaart meteen waarom Google-Extended blokkeren zo weinig oplevert: de toegang die voor AI Overviews telt loopt via Googlebot, en die staat bijna overal open.

Nieuwsmedia zijn sterk vertegenwoordigd onder de blokkeerders

Onder de sites die een AI-zoekbot blokkeren zijn nieuwsmedia en omroepen opvallend aanwezig. NRC, de Volkskrant, De Telegraaf, Trouw, Het Parool, het Financieele Dagblad, De Correspondent en De Groene Amsterdammer blokkeren de zoekbots, net als de grote regionale dagbladen, van het Brabants Dagblad en BN DeStem tot De Gelderlander, de PZC, De Limburger, de Leeuwarder Courant en Dagblad van het Noorden, plus regionale omroepen als Omrop Fryslân, Omroep West en RTV Noord. Ook NU.nl en AD, en aan overheidskant Overheid.nl, DigiD en de gemeente Amsterdam, staan in deze groep. De vier grootste nieuwsmarken laten het scherp zien: NOS, NU.nl, AD en RTL blokkeren allemaal PerplexityBot, en NU.nl en AD bovendien de zoekbot van ChatGPT.

Voor uitgevers is het blokkeren van AI-crawlers vaak een uitgesproken positie, ingegeven door auteursrecht en onderhandelingen over licenties, dus bij deze groep ligt een bewuste keuze meer voor de hand dan elders. Het meetbare gevolg blijft hetzelfde: deze titels zijn niet citeerbaar in de zoekgebaseerde antwoorden van de assistenten die ze blokkeren. Welke bronnen die assistenten dan wel aanhalen over Nederlandse onderwerpen, hebben we hier niet gemeten. Dat is de vraag van een vervolgonderzoek.

Vier grote nieuwsmerken, los gecontroleerd

Vier van de bekendste Nederlandse nieuwsmerken, NOS, NU.nl, AD en RTL, waren in de geautomatiseerde crawl onbereikbaar en zijn daarom los opgehaald en op 17 juni 2026 met de hand gecontroleerd. Ze zijn met dezelfde logica meegeteld in de cijfers en we lichten ze hier apart uit, omdat ze het patroon scherp tonen. Alle vier blokkeren GPTBot, PerplexityBot en Google-Extended. En alle vier laten Googlebot open, dus geen van vieren haalt zichzelf met het blokkeren van Google-Extended uit AI Overviews. Waar het uiteenloopt is de zoekbot van ChatGPT: NU.nl en AD, allebei van DPG Media, blokkeren ook OAI-SearchBot, terwijl NOS en RTL die openlaten en daarmee citeerbaar blijven in ChatGPT-zoek. Bij Claude is het net zo: NOS en RTL laten Claude-SearchBot open, terwijl NU.nl en AD een generieke regel voor Claude hanteren die onder standaard prefix-matching ook de huidige Claude-SearchBot meepakt. Zelfs bij deze uitgebreide blokkeerders is het beeld dus niet consequent tussen de AI-aanbieders onderling.

Dicht betekent een Disallow op de root, open betekent toegestaan. Peildatum 17 juni 2026.

Merk	GPTBot	OAI-SearchBot	PerplexityBot	Claude-SearchBot	Google-Extended	Googlebot
NOS	dicht	open	dicht	open	dicht	open
NU.nl	dicht	dicht	dicht	dicht	dicht	open
AD	dicht	dicht	dicht	dicht	dicht	open
RTL	dicht	open	dicht	open	dicht	open

NU.nl en AD zetten in hun lijst een kale regel voor Claude, en die term is een prefix van Claude-SearchBot. Onder standaard prefix-matching, zoals Google die ook toepast, blokkeert dat dus ook de huidige Claude-SearchBot. NU.nl en AD delen vrijwel exact dezelfde robots.txt, inclusief hetzelfde DPG-copyrightblok.

Verouderde blocklists

Eén bevinding wijst wel duidelijk op achterstallig onderhoud in plaats van een keuze. De verouderde Anthropic-user-agents anthropic-ai (12,7%) en Claude-Web (12,0%) worden vaker geblokkeerd dan de huidige Claude-SearchBot (8,5%) en Claude-User (8,5%). Anthropic heeft die oude namen afgeschaft, dus een regel die ze blokkeert doet niets meer, terwijl de actuele Claude-crawler intussen openstaat. Voor zo'n combinatie bestaat geen zinnige strategie, en het is daarmee een meetbaar teken dat een deel van de robots.txt-bestanden dateert van voor de huidige naamgeving van de crawlers.

Wie blokkeert, blokkeert breed

Wie een AI-crawler weert, weert er meestal niet één. Van de bijna 90 sites die minstens één van de 17 getelde AI-crawlers blokkeren, schakelt ruim de helft, 51 sites, er tien of meer tegelijk uit. Gemiddeld gaat het om tien crawlers per blokkerende site en de mediaan ligt op elf, terwijl maar 16 sites er precies één blokkeren. Dat wijst op brede lijsten die in één keer worden

overgenomen in plaats van een afweging per bot. Het duidelijkste voorbeeld zijn NU.nl en AD, allebei van DPG Media, met een vrijwel identieke robots.txt tot en met hetzelfde DPG-copyrightblok bovenaan. Het is in de praktijk een gedeeld sjabloon.

Geteld over de 17 AI-crawlers, dus zonder Googlebot en Applebot. Onder de 89 sites die er minstens één blokkeren:

Aantal geblokkeerde AI-crawlers	Aantal sites
Precies 1	16
2 tot 4	9
5 tot 9	13
10 of meer	51

De nieuwe instrumenten liggen ongebruikt

De middelen die speciaal voor het sturen van AI zijn bedoeld, worden in Nederland nauwelijks gebruikt. Een llms.txt-bestand, bedoeld om AI-assistenten te wijzen op wat ze mogen gebruiken, staat op 5,9% van de leesbare sites. De nosnippet-richtlijn, waarmee je content uit AI-samenvattingen en snippets houdt, troffen we op de gemeten pagina's niet aan. Deze twee cijfers komen uit de volledige meetronde, waarin naast robots.txt ook llms.txt en de homepage zijn opgehaald. De praktijk draait dus vrijwel volledig op robots.txt, met alle ruwheid van dien.

Wat betekent dit voor jou

1. Ken het verschil tussen trainingsbots, zoekindexbots en live-fetch-bots. Alleen de laatste twee bepalen of je in AI-antwoorden verschijnt.
2. Controleer je robots.txt op OAI-SearchBot, PerplexityBot en Claude-SearchBot. Blokkeer je die, dan ben je niet citeerbaar in AI-zoek, ook als je verder alles goed doet. Controleer of dat is wat je wilt.
3. Weet dat Google-Extended blokkeren je niet uit AI Overviews haalt. Daar kom je via robots.txt alleen uit door Googlebot te blokkeren, wat je organische verkeer kost. De nieuwe Search Console-opt-out die Google op 3 juni 2026 aankondigde, wordt voorlopig alleen in het Verenigd Koninkrijk getest en is in Nederland nog niet beschikbaar.

Cijfers

Kerncijfers, berekend over 426 leesbare domeinen.

Cijfer	Waarde
Blokkeert minstens 1 AI-crawler	20,4%
Blokkeert alleen een trainingsbot (zoekindexbots open)	6,6%
Blokkeert een AI-zoekindexbot (niet citeerbaar in AI-zoek)	13,1%
Blokkeert Google-Extended maar niet Googlebot	13,4%
Blokkeert Googlebot	0,7%
Blokkeert alleen verouderde Anthropic-UA's	0,2%
Heeft llms.txt (volledige meetronde)	5,9%

Nosnippet aangetroffen (volledige meetronde)	0,0%
--	------

Geblokkeerd per crawler, over 426 leesbare domeinen.

Crawler	Eigenaar	Categorie	Geblokkeerd (%)	Aantal
GPTBot	OpenAI	Training	17,6	75
CCBot	Common Crawl	Training	16,7	71
Amazonbot	Amazon	Gemengd	15,0	64
ClaudeBot	Anthropic	Training	14,8	63
Bytespider	ByteDance	Training	14,6	62
Google-Extended	Google	Training	14,1	60
anthropic-ai	Anthropic	Verouderd	12,7	54
ChatGPT-User	OpenAI	Live-fetch	12,4	53
PerplexityBot	Perplexity	Zoekindex	12,4	53
Applebot-Extended	Apple	Training	12,2	52
Claude-Web	Anthropic	Verouderd	12,0	51
Meta-ExternalFetcher	Meta	Live-fetch	10,1	43
Meta-ExternalAgent	Meta	Training	9,9	42
Perplexity-User	Perplexity	Live-fetch	9,4	40
OAI-SearchBot	OpenAI	Zoekindex	9,2	39
Claude-SearchBot	Anthropic	Zoekindex	8,5	36
Claude-User	Anthropic	Live-fetch	8,5	36
Applebot	Apple	Zoekindex	3,8	16
Googlebot	Google	Zoekindex	0,7	3

Methode en verantwoording

We onderzochten de 500 populairste .nl-domeinen volgens de Tranco-ranglijst. Per domein haalden we het openbare robots.txt-bestand op en bepaalden we per AI-crawler of de root volledig geblokkeerd is met een Disallow op /, met correcte groepsselectie, prefix-matching van user-agents en de regel dat een Allow wint bij een gelijke matchlengte. Er komt geen AI-model aan te pas, de meting is deterministisch en herhaalbaar.

Van de 500 domeinen waren er in de geautomatiseerde crawl 460 bereikbaar en 40 onbereikbaar. Van de 460 bereikbare hadden er 422 een leesbare robots.txt; de overige 38 sloten dat bestand af met een 403 of 401 en zijn uitgesloten in plaats van als niet-blokkerend meegeteld. Vier grote nieuwsmerken (NOS, NU.nl, AD en RTL) zaten bij de onbereikbare door de netwerkconfiguratie van de meetmachine; hun openbare robots.txt is rechtstreeks opgehaald, met exact dezelfde logica gescand en toegevoegd, wat de noemer op 426 brengt. Het robots.txt-bestand haalden we op met een browser-fingerprint, omdat veel sites

geautomatiseerde user-agents weigeren bij dit openbare bestand. De llms.txt- en nosnippet-cijfers komen uit de volledige meetronde waarin ook llms.txt en de homepage zijn opgehaald.

De crawler-indeling is geverifieerd tegen de documentatie van OpenAI, Google, Perplexity, Anthropic, Apple en Meta. We rapporteren wat een robots.txt declareert en wat daarvan volgens die documentatie het gevolg is voor de zichtbaarheid in AI. We leiden geen bedoeling af, want waarom een site een crawler blokkeert is uit het bestand niet op te maken, en een blokkade kan net zo goed een bewuste keuze als een vergissing zijn. Voor de toewijzing van regels aan crawlers gebruiken we prefix-matching, zoals Google dat ook doet: een generieke regel voor bijvoorbeeld Claude geldt daarmee ook voor de specifieke Claude-crawlers zoals Claude-SearchBot.

Drie kanttekeningen horen erbij. Een robots.txt-meting toont gedeclareerd beleid, niet of een bot zich eraan houdt. Trainings- en zoekindexbots respecteren het doorgaans, maar de live-fetch-bots zoals ChatGPT-User en Perplexity-User doen dat wisselend. En dit is een momentopname op één peildatum, want robots.txt-bestanden veranderen. De volledige dataset is op aanvraag beschikbaar voor onderzoekers en journalisten.

Bronvermelding

Bron: Timmermans Media, AI-crawleronderzoek Nederland, juni 2026, timmermansmedia.nl, peildatum 17 juni 2026.