TECHNISCH ONDERZOEKSRAPPORT · GEO & E-COMMERCE

AI-zichtbaarheid meten: één ChatGPT-check schiet tekort

Kai Heinink

MarketingCollega B.V., Enschede, Nederland

Een benchmark van 342 ChatGPT-antwoorden, 30 open koopcases, 24 exacte producten, 50 op Nederland gerichte webshops en 20 gematchte paginaparen.

Rapport	MC-R-2026-01	**Versie**	1.0
Onderzoeksdatum	4 september 2026	**Publicatie**	5 september 2026

Abstract

Op 4 september 2026 onderzocht MarketingCollega met web-enabled ChatGPT Work Mode-agents hoe herhaalbaar product- en webshopkeuzes waren bij Nederlandstalige koopvragen. Dertig open koopcases zijn elk uitgevoerd in drie afzonderlijk getaskte runs van drie beurten, samen 270 antwoorden. Slechts vier van de dertig cases leverden driemaal dezelfde webshopwinnaar op. Dit resultaat omvat drie pilotcases; wanneer die volledig worden weggelaten, waren drie van de 27 cases stabiel. In een tweede module stond bij 24 producten de exacte uitvoering inclusief EAN/GTIN vooraf vast. Bij twaalf van de 24 producten wisselde de gekozen webshop tussen de drie runs.

De geselecteerde webshop was in 61 van 71 numeriek vergelijkbare antwoorden de goedkoopste of gedeeld goedkoopste binnen de maximaal drie aanbieders die het antwoord zelf noemde. In alle negentig uiteindelijke webshopkeuzeturnen stond minimaal één URL van het domein van de gekozen webshop. Daarnaast is bij vijftig op Nederland gerichte shops één gekozen productpagina per domein technisch en inhoudelijk geauditeerd. Product JSON-LD werd daar vaker als aanwezig gecodeerd dan machineleesbare variant-, verzend- en retourinformatie. In een gematchte vergelijking van twintig paren - één aanbevolen en één niet-genoemde productpagina voor hetzelfde product - bleef echter geen van zeventien testbare auditvelden statistisch onderscheidend na Holm-correctie.

De hoofdconclusie is dat AI-zichtbaarheid binnen een vaste testset als variatie en herhaalbaarheid moet worden gerapporteerd, niet als één positie. De studie is exploratief, niet peer-reviewed en niet representatief voor alle Nederlandse webshops of ChatGPT-gebruikers. Omdat de hercodeerbetrouwbaarheid van verschillende technische velden onder de vooraf gekozen veldspecifieke drempel bleef, zijn de technische veldresultaten nadrukkelijk exploratief.

Trefwoorden: AI-zichtbaarheid, generative engine optimization, GEO, e-commerce, ChatGPT, webshops, herhaalbaarheid.

> **Exploratief onderzoeksrapport · niet peer-reviewed · dagsnapshot van 4 september 2026**

Kernresultaten

26 van de 30 open koopvragen hadden geen vaste webshopwinnaar. Slechts vier cases leverden in alle drie runs dezelfde genormaliseerde winnaar op.

4/30
open koopcases met driemaal dezelfde winnaar

12/24
exacte SKU's met driemaal dezelfde winnaar

61/71
laagste of gedeeld laagste binnen de genoemde shortlist

90/90
keuzes met minimaal één URL van het winnaar-domein

Onderdeel	Omvang
Open koopmodule	30 cases × 3 runs × 3 beurten = 270 antwoorden
Exacte-SKU-module	24 producten × 3 runs = 72 antwoorden
Webshopaudit	50 shops; één gekozen productpagina per domein
Gematchte pagina-audit	20 paren bij selectie; 14 in de strikte auditset

De resultaten beschrijven één onderzoeksopzet op één dag met web-enabled ChatGPT Work Mode-agents. Zij zijn geen marktbrede kansen, geen test van native ChatGPT Shopping en geen bewijs voor causale rankingfactoren.

Inhoud

Waarom dit onderzoek nodig is

Generative Engine Optimization, meestal afgekort tot GEO, gaat over zichtbaarheid in antwoorden van generatieve zoek- en antwoordsystemen. De term werd in academisch onderzoek geformaliseerd als een manier om zichtbaarheid in generatieve antwoorden systematisch te definiëren en te optimaliseren.[1]

Voor webshops bestaat “zichtbaarheid” echter uit meerdere beslissingen. Een systeem kan:

1. een product of productfamilie noemen;
2. de juiste uitvoering herkennen;
3. een webshop in de shortlist opnemen;
4. één aanbieder als beste keuze aanwijzen;
5. informatie van de webshop als bron gebruiken.

Die stappen worden in veel GEO-metingen op één hoop gegooid. Een merk kan bijvoorbeeld vaak worden genoemd, terwijl de aankoop steeds naar een andere webshop wordt gestuurd. Andersom kan een webshop voor één exacte SKU worden gekozen zonder breed zichtbaar te zijn binnen de categorie.

ChatGPT kan actuele webbronnen raadplegen en antwoorden van bronverwijzingen voorzien. OpenAI waarschuwt zelf dat zoekresultaten en citaties onvolledig, verouderd of onjuist kunnen zijn en dat plaatsing niet is gegarandeerd.[2] In de officiële uitleg over native shoppingresultaten noemt OpenAI onder andere intentie, productmetadata, beschikbaarheid, prijs, kwaliteit en de rol van de verkoper.[3] Ons onderzoek mat niet die native Shopping-interface, maar deze documentatie maakt wel duidelijk waarom product- en merchantdata afzonderlijk moeten worden bekeken.

Onderzoeksvraag

> **Hoe herhaalbaar zijn product- en webshopkeuzes in web-enabled ChatGPT-antwoorden voor Nederlandse koopvragen, en welke observeerbare webshop- en productpagina-eigenschappen komen in deze benchmark met die keuzes samen voor?**

Deelvragen

- Hoe vaak leveren drie uitvoeringen van dezelfde open koopcase dezelfde webshopwinnaar op?
- Hoeveel stabieler of instabieler is de merchantkeuze wanneer een exacte SKU vooraf vaststaat?
- Hoe vaak is de gekozen webshop de goedkoopste van de in het antwoord genoemde aanbiedingen?
- Welke soorten websites worden als bron gebruikt in de uiteindelijke webshopkeuze?
- Hoe compleet zijn toegang, structured data en zichtbare commerciële informatie bij vijftig op Nederland gerichte shops?
- Verschillen aanbevolen productpagina’s observeerbaar van vergelijkbare, niet-genoemde pagina’s?

Onderzoeksopzet

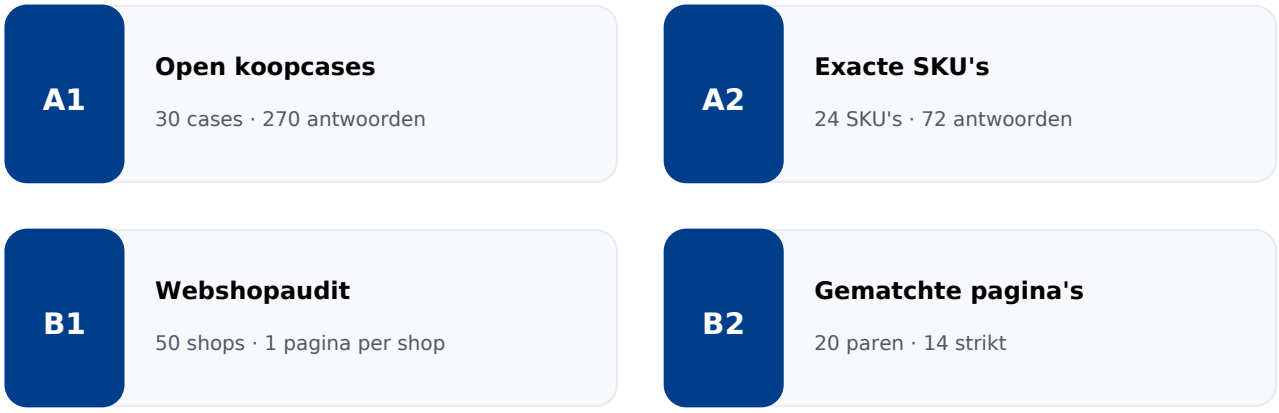
Het onderzoek bestond uit vier modules. Productkeuze, webshopkeuze, domeinkwaliteit en paginaverschillen zijn bewust niet als één meting behandeld.

Protocolstatus en registratie

Het interne onderzoeksprotocol, de vaste promptsets, het shop-panel en de analysevelden zijn vóór het hoofdveldwerk vastgelegd en als versie 1.0 bewaard. De studie is niet vooraf geregistreerd in een extern openbaar onderzoeksregister. Drie pilotcases zijn vóór de hoofdmeting uitgevoerd en later in de primaire A1-analyse behouden; daarom rapporteren we aanvullend een sensitiviteitsanalyse waarin deze cases volledig zijn weggelaten. Waar deze publicatie spreekt over een vooraf gekozen methode of drempel, bedoelen we het intern bevroren protocol en niet een extern gepreregistreerde studie.

De beschrijvende statistieken, Jaccard-overlap, betrouwbaarheidsberekeningen, sensitiviteitsanalyses en toetsen zijn uitgevoerd met eigen JavaScript/Node-analysescripts. Prompts, protocolbestanden, geaggregeerde uitkomsten en gebruikte scripts zijn opgenomen in de downloadbare onderzoeksbestanden.

Module	Wat is getest?	Omvang	Primaire uitkomst
A1 · Open koopcases	Van behoefte naar product en webshop	30 cases × 3 runs × 3 beurten	Stabiele webshopwinnaar en top-3-overlap
A2 · Exacte SKU’s	Webshopkeuze voor een vast product en EAN/GTIN	24 SKU’s × 3 runs	Stabiele webshopwinnaar en prijspositie
B1 · Webshopaudit	Toegang, productdata, commerciële informatie en vertrouwen	50 shops	Aanwezig, afwezig of onbekend per veld
B2 · Pagina-audit	Aanbevolen pagina tegenover een vergelijkbare controlepagina	20 geldige paren	Verschillen per technisch en zichtbaar veld



Figuur M1. Van vraag naar analyse.

Module A1: dertig open koopcases

De dertig vooraf geformuleerde scenario’s waren verdeeld over tien categorieën, waaronder elektronica, wonen, mode, sport, fietsen, beauty, speelgoed en huisdieren. De vraag noemde een gebruikssituatie, criteria, een budget en Nederland als markt, maar geen merk.

Iedere run bestond uit drie vaste beurten:

1. noem drie concrete producten;
2. vergelijk die producten en kies er exact één;
3. vergelijk maximaal drie webshops en kies er exact één.

De derde beurt vroeg expliciet om totale prijs, voorraad, bezorging, retourtermijn en betrouwbaarheid mee te nemen. De dertig scenario’s zijn ieder drie keer via afzonderlijk getaskte agentruns uitgevoerd. Dit leverde negentig volledige koopreizen en 270 antwoorden op.

Module A2: 24 vooraf vastgelegde producten

In de tweede module stond de productkeuze niet meer open. Er zijn 24 merkproducten uit acht categorieën geselecteerd. Van elk product waren model, uitvoering, MPN waar beschikbaar en EAN/GTIN vooraf bevroren. Ieder product was bij de selectie aantoonbaar verkrijgbaar bij minimaal vier op Nederland gerichte webshops.

De vraag was steeds: bij welke webshop kan dit exacte product nu het beste worden gekocht? Het antwoord mocht maximaal drie aanbieders noemen en moest één winnaar aanwijzen. Iedere SKU is drie keer getest. Dat leverde 72 antwoorden op.

Module B1: audit van vijftig webshops

Vijftig op Nederland gerichte shops zijn doelgericht geselecteerd in tien winkelcategorieën, met vijf shops per categorie. Per shop is één toegankelijke productpagina beoordeeld, aangevuld met openbare robots-, sitemap-, verzend-, retour- en service-informatie.

Een shop gold als op Nederland gericht wanneer deze Nederlandstalige informatie en europrijzen bood, aan Nederlandse adressen leverde en een openbaar retourbeleid had. Het bedrijf hoefde niet in Nederland gevestigd te zijn.

De audit keek onder meer naar:

- bereikbaarheid en indexeerbaarheid;
- OAI-SearchBot en ChatGPT-User;
- Product- en Offer-structured data;
- merk, identifier, prijs, voorraad en variantinformatie;
- verzend- en retourinformatie;
- productreviews, adviescontent en service-informatie.

Niet-zichtbare, geblokkeerde of niet-beoordeelbare informatie is als onbekend vastgelegd. Onbekend betekent dus niet automatisch afwezig.

Module B2: gematchte productpagina’s

Voor de 24 vaste SKU’s is geprobeerd een door ChatGPT aanbevolen verkooppagina te koppelen aan een niet-genoemde productpagina voor exact dezelfde uitvoering. Bij de selectie mocht de totale prijs maximaal tien procent verschillen.

Voor twintig SKU’s kon bij selectie een geldig paar worden samengesteld. De veertig pagina’s zijn rolblind gecodeerd: de codeur zag tijdens de veldbeoordeling niet welke pagina als aanbevolen of controle was geselecteerd, maar het merchantdomein en het paarlidmaatschap waren wel zichtbaar. Bij een latere strikte hercontrole voldeden veertien paren nog gelijktijdig aan drie voorwaarden: de exacte variant was verifieerbaar, beide aanbiedingen waren op voorraad en het prijsverschil bleef maximaal tien procent.

Wat gold als een stabiele winnaar?

Een webshop gold als winnaar wanneer het antwoord die winkel expliciet als beste keuze aanwees. Een case was volledig stabiel wanneer in alle drie runs dezelfde genormaliseerde webshop won.

Daarnaast is de overeenkomst tussen de maximaal drie genoemde webshops berekend met de gemiddelde pairwise Jaccard-index:

> **overlap = aantal webshops in beide sets ÷ aantal unieke webshops in beide sets**

Een score van 1 betekent dat twee antwoorden exact dezelfde merchantset noemden. Een score van 0 betekent dat er geen overlap was. De gerapporteerde score is het gemiddelde van de drie paarsgewijze vergelijkingen binnen een case.

Resultaat 1: bij 26 van de 30 open koopvragen wisselde de winnaar

In vier van de dertig open koopcases won bij alle drie de uitvoeringen dezelfde webshop. In 26 cases veranderde de winnaar minimaal één keer. De gemiddelde overlap van de genoemde webshop-top drieën was 0,296 op een schaal van 0 tot 1.

Uitkomst open koopcases	Resultaat
Cases met drie keer dezelfde webshopwinnaar	4 van 30
Cases met minimaal één andere winnaar	26 van 30
Gemiddelde top-3-overlap van webshops	0,296
Aantal volledige koopreisruns	90
Aantal antwoorden	270

Figuur 1. Dezelfde koopvraag, vaak een andere webshopwinnaar.

Controle zonder pilotcases

Negen pilotruns voor drie scenario’s zijn in de volledige A1-dataset opgenomen. Wanneer die drie cases volledig worden weggelaten, blijft het patroon vrijwel gelijk: drie van de 27 cases hadden drie keer dezelfde webshopwinnaar. De gemiddelde top-3-overlap was dan 0,298 tegenover 0,296 in de volledige set.

Wat dit laat zien

Een eenmalige controle kan toevallig een winnaar vastleggen die bij een nieuwe uitvoering niet terugkomt. Een nulmeting hoort daarom meerdere runs per vraag te bevatten.

Wat dit niet bewijst

Vier van de dertig is geen algemene kans dat ChatGPT een webshop stabiel kiest. De scenario’s waren doelgericht samengesteld, op één dag uitgevoerd en draaiden niet in de reguliere consumenteninterface.

Resultaat 2: ook bij een exact product wisselde de webshop in de helft van de cases

Wanneer merk, model, uitvoering en EAN/GTIN vooraf vaststonden, was de merchantkeuze vaker volledig stabiel. Bij twaalf van de 24 SKU’s won in alle drie runs dezelfde webshop. Bij de overige twaalf wisselde de winnaar minimaal één keer.

De gemiddelde top-3-overlap was descriptief 0,296 in de open koopmodule en 0,440 in de exacte-SKU-module. In de 72 antwoorden werden in totaal 196 webshopvermeldingen gecodeerd, verdeeld over 74 genormaliseerde merchantnamen.

Uitkomst exacte SKU’s	Resultaat
SKU’s met drie keer dezelfde webshopwinnaar	12 van 24
SKU’s met minimaal één andere winnaar	12 van 24
Gemiddelde top-3-overlap van webshops	0,440
Aantal antwoorden	72

Descriptieve vergelijking van verschillend samengestelde modules; geen effectschatting.
Figuur 2. Volledig stabiele webshopwinnaars per onderzoeksprotocol.

Wat dit laat zien

Productzichtbaarheid en merchantzichtbaarheid zijn verschillende KPI’s. Zelfs wanneer het product niet meer gekozen hoeft te worden, staat de verkoper niet automatisch vast.

Wat dit niet bewijst

A1 en A2 hadden andere prompts, categorieën en productselecties. Het verschil mag niet causaal aan het opnemen van een EAN of aan één afzonderlijk kenmerk worden toegeschreven.

Productidentiteit verdient een eigen controle

In de primaire codering bleef in 72 van de 72 exacte-SKU-antwoorden de vooraf vastgelegde productidentiteit behouden. Bij de open koopreisruns werd in 82 van de negentig webshopkeuzes dezelfde gecodeerde uitvoering meegenomen als in de productkeuze; vier keer was dat niet zo en vier keer bleef het onduidelijk.

De A2-productidentiteit, merchantselectie en prijspositie zijn niet onafhankelijk dubbel gecodeerd. Dit is dus geen algemene nauwkeurigheidsscore voor ChatGPT. De 24 SKU’s waren juist geselecteerd op heldere identifiers en brede verkrijgbaarheid, en niet iedere productclaim is onafhankelijk geverifieerd. Het resultaat laat vooral zien waarom merk, model, GTIN/EAN, MPN, kleur, maat, capaciteit en bundel vóór de merchantvergelijking moeten worden vastgezet.

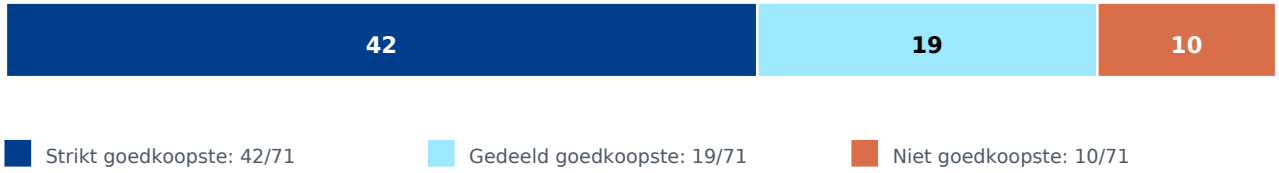
Resultaat 3: de winnaar was vaak de goedkoopste genoemde aanbieder, maar niet altijd

Van de 72 exacte-SKU-antwoorden konden er 71 numeriek op prijs worden vergeleken. In 42 antwoorden won de strikt goedkoopste genoemde aanbieder. Negentien keer was de winnaar gedeeld goedkoopste. In tien antwoorden koos ChatGPT een duurdere webshop uit de eigen shortlist.

Positie van de winnaar binnen de genoemde aanbiedingen	Aantal	Aandeel van 71
Strikt goedkoopste	42	59,2%
Gedeeld goedkoopste	19	26,8%
Niet goedkoopste	10	14,1%

Door afronding op één decimaal tellen de deelpercentages op tot 100,1 procent.

Bij de tien duurdere keuzes bedroeg de mediane meerprijs ten opzichte van de goedkoopste genoemde optie €5,87, oftewel 2,3 procent. De grootste gecodeerde meerprijs was €21,74, oftewel 29,3 procent.



Strikt goedkoopste: 42/71 Gedeeld goedkoopste: 19/71 Niet goedkoopste: 10/71

In 1 aanvullend antwoord was een numerieke prijsvergelijking niet mogelijk.

Figuur 3. Was de gekozen webshop de goedkoopste genoemde aanbieder?
Verdeling binnen 71 numeriek vergelijkbare antwoorden: 42 strikt goedkoopst, 19 gedeeld goedkoopst en 10 niet goedkoopst. Eén aanvullend antwoord was niet numeriek vergelijkbaar.

Wat dit laat zien

Prijs viel vaak samen met de uiteindelijke keuze, maar de antwoorden functioneerden niet als een zuivere prijsranglijst.

Wat dit niet bewijst

We vergeleken alleen de maximaal drie aanbiedingen die het antwoord zelf noemde, niet de volledige markt. De prompt vroeg bovendien ook om voorraad, levering, retouren en betrouwbaarheid mee te wegen. Uit deze analyse kan daarom niet worden afgeleid dat prijs de belangrijkste selectieoorzaak was.

Resultaat 4: genoemd worden en gekozen worden zijn twee verschillende dingen

In de open koopmodule kwamen 104 verschillende genormaliseerde merchantnamen voor. Daarvan wonnen er 45 minimaal één run. In de exacte-SKU-module werden 74 merchants genoemd en wonnen er 24 minimaal één keer.

Deze aantallen maken zichtbaar waarom mention share en win share niet mogen worden samengevoegd. We publiceren in het hoofdrapport bewust geen ranglijst per merchant: assortiment, categorie, prijs, beschikbaarheid en het aantal relevante verkoopkansen verschilden sterk. Er bestaat geen gelijke exposure-denominator waarmee shops eerlijk kunnen worden vergeleken. De volledige merchanttabel staat in het werkboek en de onderzoeksbijlage.

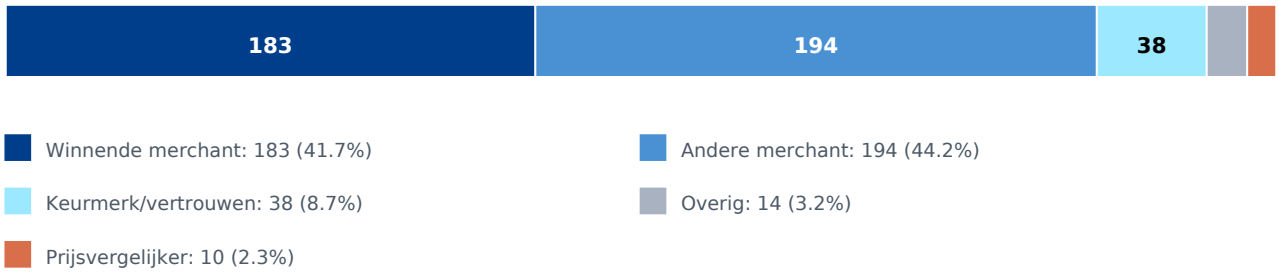
Resultaat 5: de gekozen webshop was steeds onderdeel van het bronnenbeeld

In de webshopgerichte derde beurt telden we 439 URL-vermeldingen. Daarvan verwezen er 183 naar het domein van de winnende merchant en 194 naar een andere genoemde merchant. Samen waren merchantdomeinen goed voor 377 van de 439 URL's, oftewel 85,9 procent.

Type bron in de webshopkeuzefase	URL-vermeldingen	Aandeel
Winnende merchant	183	41,7%
Andere genoemde merchant	194	44,2%
Keurmerk of vertrouwensbron	38	8,7%
Overig	14	3,2%
Prijsvergelijker	10	2,3%
Totaal	**439**	**100%**

Door afronding op één decimaal tellen de deelpercentages op tot 100,1 procent.

Alle negentig uiteindelijke webshopkeuzes bevatten minimaal één URL van het domein van de gekozen webshop.



Merchantdomeinen samen: 377/439 URL-vermeldingen (85,9%).

Figuur 4. Bronmix in de uiteindelijke webshopkeuze.

Wat dit laat zien

Iedere webshopkeuzeturn bevatte minimaal één URL op het domein van de winnaar. Dat is URL-co-occurrence; er is niet vastgesteld dat iedere URL een specifieke product- of commerciële claim ondersteunde. Externe keurmerken, publishers en vergelijkers vormden een kleiner aanvullend deel van de getelde URL's.

Wat dit niet bewijst

Een URL naast een antwoord bewijst niet dat die pagina een specifieke claim heeft veroorzaakt of volledig ondersteunt. De teleenheid is bovendien een URL-vermelding; meerdere pagina's van hetzelfde domein kunnen meetellen. De onderzoeksprompt vroeg expliciet om actuele commerciële kenmerken, waardoor merchantpagina's waarschijnlijk extra relevant waren.

Resultaat 6: in de auditsteekproef werd basisproductdata vaker als aanwezig gecodeerd

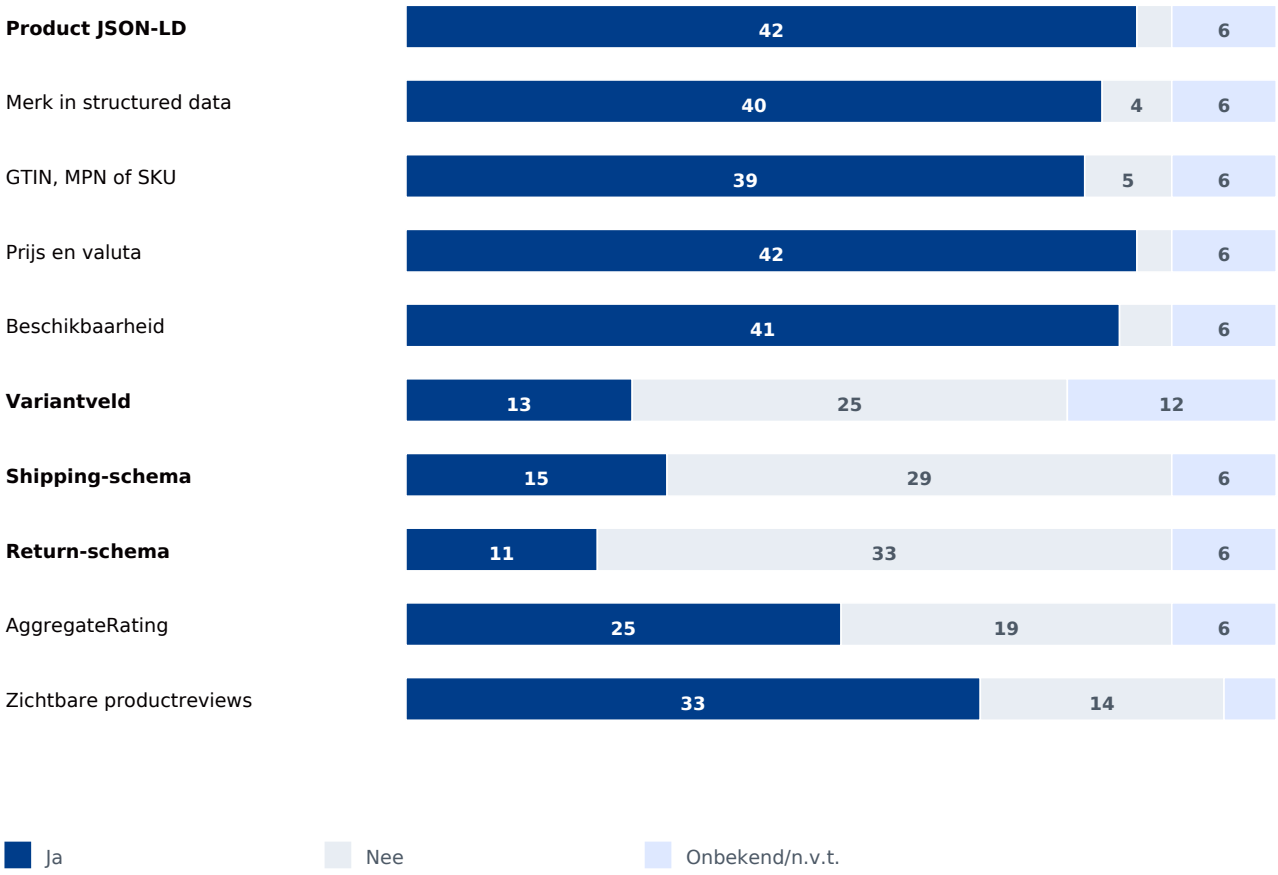
Op 42 van de vijftig onderzochte productpagina’s werd Product JSON-LD aangetroffen. Bij twee pagina’s werd het als afwezig gecodeerd; bij zes kon dit niet betrouwbaar worden vastgesteld. Velden voor merk, identifier, prijs en voorraad waren eveneens vaak aanwezig wanneer de pagina beoordeelbaar was.

In de primaire audit werden variant-, verzend- en retourvelden minder vaak als aanwezig gecodeerd. De hercodeerbetrouwbaarheid van juist deze velden bleef onder de vooraf gestelde drempel; de exacte aantallen zijn daarom exploratief:

Gecontroleerd veld	Ja	Nee	Onbekend of n.v.t.
Product JSON-LD	42	2	6
Merk in structured data	40	4	6
GTIN, MPN of SKU	39	5	6
Prijs en valuta	42	2	6
Beschikbaarheid	41	3	6
Variantveld	13	25	12
Shipping-schema	15	29	6
Return-schema	11	33	6
AggregateRating	25	19	6
Zichtbare productreviews	33	14	3

Van de 42 pagina’s met Product JSON-LD hadden er negen zowel machineleesbare verzend- als retourinformatie.

De openbare robotsregels werden bij 47 domeinen gecodeerd als toestaan van OAI-SearchBot; bij drie bleef dit onduidelijk of geblokkeerd. Alle vijftig shops waren via de gebruikte ChatGPT-User-readingroute bereikbaar; in de afzonderlijke HTTP-classificatie waren 49 shops bereikbaar en werd één shop als geblokkeerd gecodeerd. Deze observaties zijn geen garantie dat ieder toekomstig ChatGPT-User-verzoek of iedere zoekcrawl slaagt. OpenAI gebruikt OAI-SearchBot voor geschiktheid voor ChatGPT Search; toestaan is één voorwaarde voor die zoekroute, maar bewijst geen crawl, indexatie, vertoning of aanbeveling.[4]



Figuur 5. Aanwezige en ontbrekende productinformatie.
Verdeling van ja, nee en onbekend of niet van toepassing voor de gecontroleerde velden op vijftig productpagina’s.

Wat dit laat zien

In deze primaire audit waren basisvelden vaker aanwezig gecodeerd dan variant-, verzend- en retourvelden. Die laatste velden zijn kandidaten voor een handmatige templatecontrole; de audit bewijst niet dat dit de grootste tekortkomingen van het volledige panel zijn.

Wat dit niet bewijst

De vijftig shops vormen geen representatieve steekproef van alle Nederlandse webshops. Eén productpagina per domein beschrijft bovendien niet automatisch ieder template of iedere SKU. Aanwezigheid van structured data is ook geen bewijs dat het veld is gebruikt of een hogere selectiepositie veroorzaakt.

Waarom deze velden toch relevant zijn

OpenAI’s actuele productspecificatie vraagt voor directe feeds onder meer om titel, beschrijving, URL, merk, verkoper, afbeelding, prijs en beschikbaarheid en ondersteunt daarnaast varianten, GTIN, verzending, retouren en reviews.[5] Schema.org biedt daarnaast expliciete typen voor Product, Offer, OfferShippingDetails, MerchantReturnPolicy en AggregateRating.[6-10] Dat toont welke commercegegevens machineleesbaar kunnen worden aangeleverd. Het is geen lijst met bewezen rankingfactoren.

Resultaat 7: geen auditveld bleef statistisch onderscheidend na Holm-correctie

Voor twintig producten vergeleken we een aanbevolen verkooppagina met een niet-genoemde verkooppagina voor hetzelfde product. De pagina’s waren bij selectie maximaal tien procent in totale prijs van elkaar verwijderd. **Geen van de zeventien testbare velden bleef statistisch onderscheidend na correctie voor meervoudig toetsen.**

De vergelijking gebruikte per veld een tweezijdige exacte McNemar-toets met α = 0,05.[11] Onbekende of niet-toepasselijke waarden werden voor dat veld buiten de bekende paren gehouden. De p-waarden van de volledige familie van zeventien testbare velden zijn met de Holm-methode gecorrigeerd.[12]

In de primaire tabel kwamen shipping- en retourschema bij acht paren alleen op de aanbevolen pagina voor en bij één paar alleen op de controlepagina. Negen aanbevolen pagina’s waren echter afkomstig van bol en deelden grotendeels hetzelfde template. Binnen de bekende paren met een andere aanbevolen merchant was het netto verschil nul. De Holm-gecorrigeerde p-waarde bedroeg 0,664. In de strikte set van veertien paren was de richting vijf tegenover één, maar ook daar was het resultaat niet onderscheidend; de ongecorrigeerde p-waarde was 0,219.

De hercodeerbetrouwbaarheid haalde voor verschillende B2-velden de vooraf vastgelegde veldspecifieke drempel niet en er is geen volledige hercodering uitgevoerd. De veldvergelijkingen blijven daarom exploratief, ook waar de rekenkundige toets correct is uitgevoerd. De complete vergelijking van alle velden, inclusief de ongecorrigeerde waarden, staat in de onderzoeksbijlage.

Figuur 6. Geen los auditveld bleef onderscheidend na correctie.

Wat dit laat zien

Shipping- en retourschema blijven mogelijke hypothesen voor vervolgonderzoek. Het primaire resultaat is het nulresultaat: deze dataset levert geen robuust bewijs voor één technisch veld waarmee een webshop de selectie kan winnen.

Wat dit niet bewijst

Het resultaat bewijst niet dat techniek irrelevant is. De analyse was klein, retrospectief, niet-gerandomiseerd en geclusterd op domeinen. Ook assortiment, merkrelaties, prijs, beschikbaarheid en autoriteit konden niet worden geïsoleerd.

Waarom correctie nodig was

Wie veel velden tegelijk test, vergroot de kans dat minstens één resultaat toevallig “significant” lijkt. De Holm-correctie verlaagt dat risico door de grens per vergelijking aan te scherpen. Daarom publiceren we niet de aantrekkelijke conclusie dat shipping- of retourschema een rankingfactor is. De data ondersteunt alleen de bescheidener conclusie dat dit een richting voor groter, vooraf vastgelegd vervolgonderzoek is.

Discussie: van één positie naar gemeten variatie

De resultaten wijzen in dezelfde praktische richting. AI-zichtbaarheid bestaat niet uit één uitkomst, maar uit een keten van selecties:

1. Is de site technisch bereikbaar?
2. Wordt het juiste product begrepen?
3. Wordt de webshop genoemd?
4. Haalt de webshop de shortlist?
5. Wordt de webshop als winnaar gekozen?
6. Worden de relevante claims met bruikbare bronnen onderbouwd?
7. Gebeurt dit opnieuw bij een herhaling?

Dat is fundamenteel anders dan één klassieke rangpositie rapporteren. Ook traditionele zoekresultaten veranderen, maar bij een gegenereerd antwoord kunnen de selectie, formulering, volgorde en bronmix binnen dezelfde taak tegelijk wisselen.

Voor e-commerce is vooral het onderscheid tussen product- en merchantzichtbaarheid belangrijk. Een fabrikant kan het product winnen terwijl een retailer de transactie verliest. Een webshop kan geregeld in de shortlist staan, maar zelden de expliciete voorkeur krijgen. En een winkel kan wel worden gekozen, maar met onjuiste of verouderde informatie over prijs, voorraad of retouren.

De juiste GEO-vraag is daarom niet:

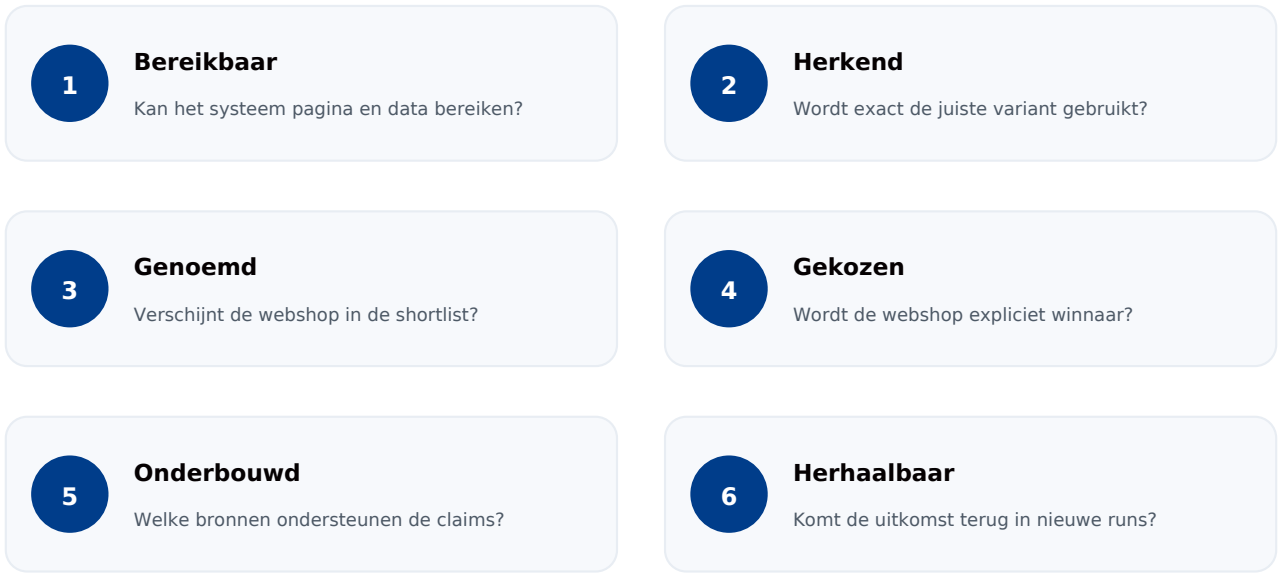
> **“Staan we in ChatGPT?”**

maar:

> **“Hoe vaak worden we voor deze concrete koopintentie genoemd en gekozen, voor welke SKU, op basis van welke bronnen en met hoeveel variatie tussen runs en meetmomenten?”**

Zes meetlagen voor AI-zichtbaarheid

Op basis van deze benchmark onderscheidt MarketingCollega zes meetlagen. Het is een praktisch analysekader, geen wetenschappelijk gevalideerde schaal of commercieel scoresysteem.



Meet de lagen afzonderlijk; tel ze niet op tot één schijnpreciese score.

Figuur 7. Zes meetlagen voor GEO bij webshops.

1. Bereikbaar

Kan het systeem de relevante pagina's en productdata technisch bereiken? Controleer onder meer indexeerbaarheid, canonical, OAI-SearchBot, serverblokkades, renderbaarheid en eventuele productfeeds.

KPI's: aandeel bereikbare SKU-pagina's, aandeel toegestane zoekcrawlerpaden en feeddekking.

2. Herkend

Wordt exact het juiste product of de juiste variant gebruikt? Leg merk, model, GTIN/EAN, MPN, kleur, maat, capaciteit, conditie en bundel vast.

KPI's: exacte-SKU-herkenning en variantfouten.

3. Genoemd

In welk aandeel van geldige antwoorden verschijnt de webshop in de shortlist?

KPI: mention share = antwoorden met merchantvermelding ÷ geldige antwoorden.

4. Gekozen

In welk aandeel van de antwoorden is de webshop de expliciete winnaar?

KPI: win share = antwoorden met merchant als winnaar ÷ geldige antwoorden.

5. Onderbouwd

Welke pagina's en domeinen worden als bron gebruikt, en ondersteunen ze daadwerkelijk prijs, voorraad, levering, retouren en verkoper?

KPI's: source coverage en afzonderlijk geverifieerde claim accuracy. Een bron-URL alleen is nog geen bewijs dat iedere claim klopt.

6. Herhaalbaar

Hoe sterk wisselen de shortlist en de winnaar wanneer dezelfde taak opnieuw wordt uitgevoerd?

KPI's: dominant-win share, top-3-Jaccard, spreiding tussen runs en verandering tussen meetgolven.

Een praktische meetopzet

Voor een webshopnulmeting adviseren we:

- werk met een vaste set koopintenties, categorieën en prioriteits-SKU's;
- splits open ontdekkingsvragen en exacte-SKU-vragen;
- voer iedere prompt meerdere keren uit;
- bewaar het volledige antwoord, alle URL's, datum, tijd en onderzoeksomgeving;
- codeer product, variant, merchants, winnaar, bronnen en commerciële claims apart;
- herhaal de meting in golven, zodat momentvariatie niet als structurele groei wordt verkocht;
- rapporteer technische auditresultaten naast, niet in plaats van, de feitelijke antwoorduitkomsten.

Dit onderzoek levert geen universeel minimumaantal runs op. Gebruik een pilot om de variatie te schatten, leg het aantal herhalingen vooraf vast en kies dit op basis van de gewenste onzekerheidsmarge, segmentatie en beslisimpact.

Praktijkadvies op basis van de benchmark en officiële specificaties

Dit onderzoek levert geen magische rankingformule op. De combinatie van de benchmark, algemene informatiehygiëne en officiële OpenAI- en Schema.org-specificaties geeft wel een logische controlevolgorde.

Prioriteit 1: maak iedere commerciële claim controleerbaar

Toon bij de concrete uitvoering:

- actuele totaalprijs;
- voorraadstatus;
- concrete bezorgbelofte;
- verzendkosten;
- retourtermijn en retourkosten;
- feitelijke verkoper bij marketplaces;
- conditie, bundel en eventuele abonnementsvoorwaarden.

Zorg dat deze informatie niet alleen in een onleesbare interface, afbeelding of na een onnodige interactie verschijnt.

Prioriteit 2: laat productidentiteit overal overeenkomen

Gebruik dezelfde merknaam, modelnaam, GTIN/EAN, MPN, variant en bundel in zichtbare tekst, structured data en feeds. Een AI-systeem dat twee uitvoeringen niet betrouwbaar uit elkaar kan houden, kan ook geen zuivere merchantvergelijking maken.

Prioriteit 3: vul structured data inhoudelijk aan

Controleer Product en Offer eerst op basisvelden. Voeg daarna waar passend variantrelaties, OfferShippingDetails, MerchantReturnPolicy en productbeoordelingen toe. Doe dit omdat volledige, consistente data beter uitleesbaar en controleerbaar is - niet omdat deze benchmark een rankingsprong bewijst.

Prioriteit 4: optimaliseer op de hele beslisketen

Categorieadvies, productdetailpagina, servicevoorwaarden, externe reviews, tests en vergelijkers kunnen verschillende rollen spelen. Een productpagina alleen hoeft niet alle informatiebehoeften op te lossen, maar de bronnen moeten elkaar niet tegenspreken.

Prioriteit 5: meet het antwoord, niet alleen de techniek

Een audit kan laten zien wat technisch ontbreekt. Alleen herhaalde antwoordmetingen laten zien of producten en merchants daadwerkelijk worden genoemd en gekozen.

Methodologische kwaliteit en betrouwbaarheid

De primaire codering is uitgevoerd door de agent die ook het antwoord of de audit produceerde. A1-keuzes en deelsets van de B1- en B2-audits zijn blind opnieuw beoordeeld. A2-winnaars, productidentiteit en prijsposities zijn niet in een volledige blinde dubbelcodering gecontroleerd.

Onderdeel	Heraudit	Overeenstemming	Cohen’s kappa[13]
A1-beschikbaarheidsclassificaties	23 runs, 53 classificaties	90,6%	0,768
B1 bekende binaire waarden	13 shops, 311 beslissingen	95,8%	0,839
B1 exacte antwoordcategorie	13 shops, 351 classificaties	88,3%	0,685
B2 bekende binaire waarden	10 pagina’s, 184 beslissingen	95,1%	0,818
B2 exacte antwoordcategorie	10 pagina’s, 230 classificaties	79,6%	0,532

Bij de hercodering van 23 A1-runs was de overeenstemming over de expliciete productwinnaar, webshopwinnaar en variantclassificatie 100 procent. Voor merchantvermeldingen was de ruwe overeenstemming 96,4 procent over 55 kandidaatclassificaties; de bijbehorende kappa is door de gebruikte kandidaatset en sterke prevalentie niet inhoudelijk interpreteerbaar.

De intern vooraf gekozen betrouwbaarheidsdrempel was $\kappa \geq 0{,}80$ voor binaire velden en $\kappa \geq 0{,}70$ voor multicategorievelden. Deze drempel werd voor verschillende afzonderlijke technische velden niet gehaald. Het protocol schreef dan volledige hercodering voor; die is niet uitgevoerd. Daarom behandelen we specifieke B1- en B2-velden als exploratief en gebruiken we ze niet om rankingfactoren te claimen.

Beperkingen

Eén meetdag

Alle runs en audits vonden plaats op 4 september 2026. Prijzen, voorraad, buyboxes, bezorgbeloften, retourvoorwaarden, zoekresultaten en modeluitkomsten kunnen daarna veranderen.

Geen reguliere consumenteninterface

De uitvoering gebruikte web-enabled ChatGPT Work Mode-agents. Modelbuild, zoekbackend, accountconfiguratie en exacte locatiecontext konden niet volledig worden vastgelegd. Dit is niet hetzelfde als de standaard ChatGPT-consumentenervaring en niet hetzelfde als native ChatGPT Shopping.

Doelgerichte steekproeven

De dertig journeys, 24 SKU’s en vijftig shops zijn bewust samengesteld om verschillende categorieën en situaties te onderzoeken. Het zijn geen aselecte steekproeven en de percentages zijn geen marktbrede prevalenties.

Afzonderlijke agenttaken zijn geen consumentensteekproef

Herhalingen van dezelfde case zijn via verschillende agenttaken uitgevoerd. Eén agent verwerkte soms meerdere andere cases achter elkaar en een volledige technische contextreset is niet voor iedere taak gelogd. We spreken daarom niet van onafhankelijke consumentensessies.

Pilotdata in de hoofdanalyse

Negen pilotruns voor drie A1-scenario’s zijn opgenomen. Bij vijf pilotruns is minimaal één eerdere beurt achteraf als reconstructed of best effort vastgelegd; de webshopkeuzebeurten waren wel verbatim bewaard. De analyse zonder deze drie cases leverde dezelfde inhoudelijke richting op.

Geen volledige claimnauwkeurigheidsaudit

Een bronvermelding, prijs of voorraadclaim in een antwoord is niet voor alle 342 antwoorden onafhankelijk tegen de live pagina geverifieerd. We publiceren daarom geen algemene nauwkeurigheidspercentages voor prijs, voorraad, levering of retouren.

A2 niet volledig dubbel gecodeerd

De winnaars, productidentiteit en prijsposities in de exacte-SKU-module komen uit de primaire codering en zijn niet voor alle 72 antwoorden blind door een tweede codeur beoordeeld. De A2-resultaten zijn daarom descriptief en niet geschikt als algemene nauwkeurigheidsscore.

Eén productpagina per domein

De B1-audit beschrijft één gekozen productpagina en openbare domeininformatie op het meetmoment. Zij vertegenwoordigt niet noodzakelijk ieder template of iedere productpagina van de shop.

Kleine, geclusterde matched analyse

De B2-analyse bevatte twintig paren en een strikte sensitiviteitsset van veertien. Negen van de twintig aanbevolen pagina’s waren van bol. Pagina’s op hetzelfde domein delen techniek en zijn geen volledig onafhankelijke observaties.

Niet alle vooraf geplande controles zijn uitgevoerd

De optionele herhaalmeting na zeven dagen vond niet plaats. Ook geplande analyses van volledige claimnauwkeurigheid, MRR, journey persistence en samengestelde paginadimensies zijn niet als bevestigende uitkomsten gepubliceerd.

Geen causaliteit of garantie

Een vermelding, bronverwijzing of winnaar in deze benchmark voorspelt geen toekomstige AI-zichtbaarheid, verkeer, conversie of omzet. De onderzoeksopzet kan samenhang en herhaalbaarheid beschrijven, maar geen causaal selectie-effect aantonen.

Protocolafwijkingen

Om de resultaten controleerbaar te houden, leggen we de belangrijkste afwijkingen van het oorspronkelijke plan expliciet vast:

- de geplande Temporary Chat-consumenteninterface is vervangen door web-enabled Work Mode-agents;
- drie pilotcases zijn in A1 opgenomen;
- bij 43 van de 81 main-runs lag volgens de registratie maximaal twee seconden tussen de eerste en derde beurt; bij negen runs waren alle drie timestamps gelijk. Deze timestamps zijn administratief en niet bruikbaar als exacte meettijden. Alleen de onderzoeksdatum wordt als betrouwbaar meetvenster gerapporteerd;
- de vooraf vastgelegde betrouwbaarheidseis is voor verschillende auditvelden niet gehaald zonder volledige hercodering;
- de strikte matched sensitiviteitsset bleef met veertien paren onder het vooraf gewenste minimum van twintig;
- de optionele tweede meetgolf na zeven dagen is niet uitgevoerd.

Conclusie

Deze benchmark maakt één fout in veel GEO-rapportages zichtbaar: een losse ChatGPT-uitkomst wordt behandeld alsof het een positie is.

Bij 26 van de dertig open koopvragen wisselde de webshopwinnaar minimaal één keer. Zelfs toen het exacte product vooraf vaststond, veranderde bij de helft van de SKU’s de winnaar tussen drie runs. De laagste genoemde prijs viel vaak samen met de selectie, merchantpagina’s domineerden de bronmix en veel shops boden al technische productdata. Toch vond de gematchte analyse geen los paginaveld dat na correctie als onderscheidend overeind bleef.

De praktische consequentie binnen deze benchmark is helder. Voor webshops ligt het voor de hand om product- en commerciële data volledig, consistent en controleerbaar te maken. Wie wil weten of die inspanning samenvalt met andere antwoorduitkomsten, moet vervolgens herhaald meten wat er gebeurt: per intentie, per SKU, per merchant en over de tijd.

AI-zichtbaarheid is niet “aan” omdat een webshop één keer verschijnt. Zij wordt pas meetbaar wanneer bereikbaarheid, herkenning, vermelding, selectie, onderbouwing en herhaalbaarheid afzonderlijk worden gevolgd.

Data en methode

De volgende onderzoeksbestanden zijn versieerbaar bewaard:

- onderzoeksprotocol en vaste prompts;
- lijst met dertig open koopcases;
- bevroren lijst met 24 exacte SKU’s;
- ruwe en gecodeerde antwoorddata;
- auditdata van vijftig shops;
- de twintig gematchte paginaparen;
- codeboek, betrouwbaarheidscontroles en analysescripts.

Onderzoeksdata AI-zichtbaarheid webshops 2026
Bestandsformaat: XLSX · Versie 1.0 · Onderzoeksdatum 4 september 2026
SHA-256 van het aangeleverde v1.0-werkboek:
18401b9bfdc1d5e8121172708b7a7bd1ae498b0e49003708a4b702fd9a4d5a49
Bestand: Download onderzoeksdata

Volledig onderzoeksrapport
Bestandsformaat: doorzoekbare PDF · Versie 1.0 · Rapport MC-R-2026-01
Bestand: Download het onderzoeksrapport

Protocol, ruwe antwoorden, codering en analysescripts
Bestandsformaat: ZIP · Versie 1.0
SHA-256: 9cbad33c0a475033f053550f7eacc1ff9044eadab7394fdf75f68bb0efd8d6fe
Bestand: Download het reproduceerbaarheidsmateriaal

Onderzoeksbijlagen in HTML
Register van koopcases, exacte producten, shops, gematchte paren, codeboek en aanvullende analyses: Bekijk de onderzoeksbijlagen

De gepubliceerde bestanden maken de methode, prompts, primaire antwoorden, codering en berekeningen controleerbaar en maken een procedurele herhaling mogelijk. Volledige computationele reproductie is niet mogelijk, omdat modelbuild, zoekbackend, sessiestatus en dynamische paginaversies niet zijn bevroren. Een nieuwe uitvoering hoeft daardoor niet dezelfde uitkomsten op te leveren.

Voorkeursbronvermelding

> **Heinink, K. (2026). AI-zichtbaarheid meten: één ChatGPT-check schiet tekort. MarketingCollega Research Report MC-R-2026-01, versie 1.0. https://www.marketingcollega.nl/onderzoeken/ai-zichtbaarheid-meten/**

Cijfers en grafieken mogen met duidelijke bronvermelding en een link naar dit onderzoek worden overgenomen.

Onderzoeksintegriteit

Uitvoerder en AI-gebruik

MarketingCollega heeft de onderzoeksvragen, afbakening en publicatie geïnitieerd. ChatGPT-agents waren zowel onderzoeksobject als onderzoeksinstrument: zij voerden de vaste taken uit, raadpleegden openbare webbronnen en verzorgden de primaire codering. Afzonderlijke agentcontroles, hercoderingen en analysescripts zijn gebruikt om uitkomsten te controleren. MarketingCollega blijft verantwoordelijk voor de redactionele interpretatie en publicatie.

Belangenverklaring

MarketingCollega biedt commerciële dienstverlening rond GEO, SEO en e-commerce. Er bestaat daardoor een relevant commercieel belang bij het onderzoeksterrein en bij positionering als specialist. Dit onderzoek identificeert geen gegarandeerde rankingfactor en doet geen belofte over toekomstige zichtbaarheid of omzet.

Geen van de onderzochte webshops, platforms of leveranciers heeft voor opname betaald of invloed gehad op de onderzoeksopzet, codering, analyse of publicatie. Eventuele overige commerciële relaties met onderzochte partijen zijn niet als selectiecriterium gebruikt en zijn in deze benchmark niet als analysevariabele onderzocht.

Financiering

Dit onderzoek is door MarketingCollega zelf gefinancierd. Er was geen externe onderzoeksfinancier of opdrachtgever.

Correcties en updates

We corrigeren aantoonbare fouten zichtbaar en overschrijven eerdere conclusies niet stilzwijgend. Iedere inhoudelijke wijziging krijgt een datum, versienummer en korte toelichting.

Versie	Datum	Wijziging
1.0	5 september 2026	Eerste openbare publicatie

References

1. Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 5-16). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671900
2. OpenAI. (z.d.). Searching the web with ChatGPT. Geraadpleegd op 4 september 2026. https://help.openai.com/en/articles/9237897-chatgpt-search
3. OpenAI. (z.d.). Shopping with ChatGPT Search. Geraadpleegd op 4 september 2026. https://help.openai.com/en/articles/11128490-shopping-with-chatgpt-search
4. OpenAI. (z.d.). Overview of OpenAI Crawlers. Geraadpleegd op 4 september 2026. https://developers.openai.com/api/docs/bots
5. OpenAI. (z.d.). Products - Agentic Commerce. Geraadpleegd op 4 september 2026. https://developers.openai.com/commerce/specs/file-upload/products
6. Schema.org. (z.d.). Product. Geraadpleegd op 4 september 2026. https://schema.org/Product
7. Schema.org. (z.d.). Offer. Geraadpleegd op 4 september 2026. https://schema.org/Offer
8. Schema.org. (z.d.). OfferShippingDetails. Geraadpleegd op 4 september 2026. https://schema.org/OfferShippingDetails
9. Schema.org. (z.d.). MerchantReturnPolicy. Geraadpleegd op 4 september 2026. https://schema.org/MerchantReturnPolicy
10. Schema.org. (z.d.). AggregateRating. Geraadpleegd op 4 september 2026. https://schema.org/AggregateRating
11. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157. https://doi.org/10.1007/BF02295996
12. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65-70.
13. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104