timmermans media.

AI-antwoorden over Nederlandse merken, getoetst tegen geverifieerde feiten

1917 antwoorden van ChatGPT, Gemini en Perplexity over 37 Nederlandse merken, beoordeeld tegen op de peildatum geverifieerde feiten. Fouten zijn zeldzaam en vooral verouderd, niet verzonnen, en het scheelt een factor 8 welk AI-model je klant gebruikt.

Auteur: Matt Timmermans, Timmermans Media

Publicatiedatum: 8 juli 2026

Metingen uitgevoerd op 5 en 6 juli 2026; grondwaarheden geverifieerd op peildatum 4 juli 2026 en waar nodig herzien op 7 juli 2026; gepreregistreerd meetvenster 3 t/m 10 juli 2026

Versie: 1.2

Versiehistorie: 1.0 (8 juli 2026) eerste uitgave; 1.1 (8 juli 2026) correcties wijzigingslog-merken (O1-062, O1-074, O1-142), tabel 4.5 (ChatGPT strik op vraagniveau), interval 4.1, en tekstuele en notatiecorrecties; 1.2 (8 juli 2026) bron O1-045 naar kpn.com en O1-046 naar vodafone.nl, ook doorgevoerd in tab 8 van het werkboek

Citeer als: Timmermans, M. (2026). AI-antwoorden over Nederlandse merken: 1917 antwoorden van ChatGPT, Gemini en Perplexity beoordeeld tegen geverifieerde feiten. Timmermans Media. Meting 5 en 6 juli 2026, peildatum grondwaarheden 4 juli 2026. timmermansmedia.nl/onderzoek/ai-hallucinaties-nederlandse-merken/

Dit document is het citeerbare naslagwerk bij de onderzoekspagina. Het bewijst de claims van die pagina en introduceert geen nieuwe analyses. Doelgroep: journalisten, onderzoekers en kritische vakgenoten.

Inhoudsopgave

1. Samenvatting

Wij stelden ChatGPT, Gemini en Perplexity 185 feitelijke klantvragen over 37 Nederlandse merken, plus 14 strikvragen met 14 gespiegelde controlevragen. Elke vraag ging drie keer naar elk model via de API in de standaard consumentenconfiguratie. Dat leverde 1917 beoordeelde antwoorden op: 1665 feitelijke antwoorden en 252 strik- en controle-antwoorden. Elk antwoord is beoordeeld tegen een op de peildatum geverifieerd feit. De primaire analyse-eenheid is het vraagniveau: het meerderheidsoordeel van de drie runs, vooraf geregistreerd. Runniveau staat er alleen bij als gelabelde sensitiviteit of waar een cijfer van nature op runniveau ligt, zoals de fouttypes en de bronnen.

De vijf dragende bevindingen, elk als citeerbare zin:

1. Op het vraagniveau geeft Perplexity bij 9,2% van de klantvragen in meerderheid een fout antwoord, ChatGPT bij 2,7% en Gemini bij 1,1%; tussen het slechtste en het beste model zit ongeveer een factor 8.
2. Van alle foute antwoorden is 62% verouderde informatie en slechts 11% echt verzonnen: de modellen hallucineren zelden, maar serveren vooral de werkelijkheid van gisteren.
3. Verzonnen fouten komen in deze meting uitsluitend bij Perplexity voor, in negen van de 1665 feitelijke antwoorden, verdeeld over zes vragen; ChatGPT en Gemini produceerden geen enkele verzonnen fout.
4. Tijdgevoelige feiten zoals prijzen en voorwaarden gaan vaker fout dan stabiele feiten: 6,8% van de tijdgevoelige vraag-modelcombinaties tegen 2,5% van de stabiele.
5. In foute antwoorden verschuift het profiel van de aangehaalde bronnen richting vergelijkers en affiliatesites en weg van het eigen merkdomein; dat een merkdomein tussen de bronnen staat, beschermt niet tegen een fout antwoord.

De meta-bevinding raakt de betrouwbaarheid van de meetlat zelf. Twaalf van de 185 geverifieerde feiten bleken tijdens het beoordelen alweer achterhaald, hoewel ze enkele dagen eerder met de hand waren gecontroleerd. In verschillende gevallen signaleerden de zoekende modellen die wijziging eerder dan de handmatige verificatie. Dit onderzoek is op die twaalf wijzigingen gecorrigeerd, waar veel hallucinatiemetingen dat niet doen en een model dan onterecht een fout aanrekenen.

2. Aanleiding en onderzoeksvragen

Nederlanders krijgen hun antwoorden steeds vaker van een AI-assistent in plaats van een zoekresultaat. Voor merken verschuift daarmee de vraag van "sta ik goed in de zoekresultaten" naar "vertelt de assistent het juiste verhaal over mijn prijzen, voorwaarden en aanbod". Over die tweede vraag is voor de Nederlandse markt weinig systematisch gemeten. Dit onderzoek vult dat gat met een gecontroleerde meting: dezelfde klantvragen, dezelfde peildatum, drie modellen, en een menselijk geverifieerde feitenlat.

Het onderzoek is vooraf geregistreerd en bevroren voor de eerste meting, met zes hypotheses en een vaste analyse-eenheid. De hypotheses luiden letterlijk, zoals vastgelegd in het preregistratieblad van het werkboek:

- H1: elk model geeft op een substantieel deel van klantvragen over Nederlandse merken een antwoord dat op de peildatum niet klopt (kopcijfer per model met 95%-betrouwbaarheidsinterval).
- H2: het foutpercentage stijgt naarmate de merkbekendheid daalt (groot, uitdager, mkb), bij alle drie de modellen en binnen de sectordrieluiken, ook gestratificeerd naar vraagtype.

- H3: het percentage geen antwoord is bij mkb-merken duidelijk hoger dan bij grote merken (co-kopcijfer).
- H4: tijdgevoelige feiten kennen een hoger foutpercentage dan stabiele feiten; het aandeel verouderd is daar groter.
- H5: modellen bevestigen valse premisses over verzonnen producten vaker dan ze ware premisses over echte producten ontkennen (strik versus controle).
- H6 (exploratief): antwoorden zonder getoonde bron zijn vaker fout dan antwoorden met bron.

De hypotheses en beslisregels zijn voorafgaand aan de meting vastgelegd in het onderzoekswerkboek; amendementen en de grondwaarheid-wijzigingslog zijn per datum gelogd (tab 8). De uitkomsten per hypothese zijn vastgelegd op 7 juli 2026 en staan in hoofdstuk 4. De analyse-eenheid is het vraagniveau, het meerderheidsoordeel van drie runs; runniveau geldt uitsluitend als gelabelde sensitiviteitsanalyse, omdat drie runs binnen een vraag niet onafhankelijk zijn.

3. Methode

3.1 Ontwerp

De vragenset bestaat uit 185 feitelijke klantvragen over 37 Nederlandse merken. De merken zijn ingedeeld naar bekendheid in drie tiers: 22 grote merken, 8 uitdagers en 7 mkb-merken. Per merk zijn vijf vragen gesteld. Naast de feitelijke set zijn er 14 strikvragen, elk met een aantoonbaar onjuiste premisse, en 14 gespiegelde controlevragen met een ware premisse. De strikvragen meten of een model een verzonnen product of voorwaarde bevestigt; de controles meten of het model een echt bestaand aanbod juist herkent.

De vragen dekken twintig sectoren, van webshops, supermarkten en telecom tot bank, energie, reizen en maaltijdboxen. Van de 185 feitelijke vragen zijn er 79 tijdgevoelig (prijzen, voorwaarden, actueel aanbod) en 106 stabiel. De sectoren zijn waar mogelijk opgezet als drieluiken waarin een groot merk, een uitdager en een mkb-merk in dezelfde markt naast elkaar staan, zodat het bekendheidseffect binnen een sector afleesbaar is.

Elke vraag is drie keer aan elk van de drie modellen gesteld via de API, in de standaard consumentenconfiguratie zonder deep research of expliciete redeneerstanden. De modelversies waren:

Model	API-modelstring
ChatGPT	gpt-5.5-2026-04-23
Gemini	gemini-3.5-flash
Perplexity	sonar [voetnoot 1]

Voetnoot 1: gemeten via de Perplexity-API met modelstring sonar. Perplexity voert in zijn consumentenproducten de naam Sonar 2; de API-string kent geen versieaanduiding, zodat niet per meting vast te stellen is welke modelversie antwoordde.

Dat geeft 185 vragen maal 3 modellen maal 3 runs is 1665 feitelijke metingen, plus 14 strik- en 14 controlevragen maal 3 modellen maal 3 runs is 252 metingen, samen 1917. De primaire eenheid is het meerderheidsoordeel van de drie runs per vraag. Bij een vraag waar de drie runs elk een ander oordeel geven, is er geen meerderheid; die vraag valt buiten de kopcijferbasis. Per model gebeurt dat precies een keer, zodat de kopcijferbasis 184 vragen is.

3.2 Grondwaarheden

Metingen uitgevoerd op 5 en 6 juli 2026; grondwaarheden geverifieerd op peildatum 4 juli 2026 en waar nodig herzien op 7 juli 2026; gepreregistreerd meetvenster 3 t/m 10 juli 2026.

Voor elke vraag is vooraf een grondwaarheid vastgelegd: het juiste antwoord op de peildatum, geverifieerd tegen een primaire bron zoals de eigen site van het merk, de voorwaarden of een officieel register. Naast de inhoudelijke verificatie is per merk een operationele-statuscheck gedaan: bestaat het bedrijf op de peildatum nog en is het actief, gecontroleerd via de eigen site, recent nieuws, de Kamer van Koophandel en het insolventieregister. Een merk dat niet meer bestaat, vervalt of verhuist naar de strikset.

Waar de drie modellen consistent van de vastgelegde grondwaarheid afweken, is die grondwaarheid opnieuw tegen de primaire bron gecontroleerd (het hercheck-protocol). Dat leidde tot twaalf herzieningen en twee bevestigingen; die staan in paragraaf 3.5. Scoren gebeurt altijd tegen de grondwaarheid op de peildatum, nooit tegen parate kennis van de beoordelaar.

3.3 Beoordeling

Elk antwoord kreeg een score in vier categorieën. Juist: het kernantwoord komt overeen met de grondwaarheid op de peildatum. Deels: de kern klopt, maar met een relevant onjuist of ontbrekend element, bijvoorbeeld het juiste mechanisme met een oud bedrag onder voorbehoud. Fout: het kernantwoord is strijdig met de grondwaarheid of noemt stellig een onjuist feit of bedrag. Geen antwoord: het model weet het niet, weigert, of vraagt alleen om verduidelijking. Deels is nadrukkelijk geen fout; waar hieronder een cijfer voor niet-volledig-juist staat, telt deels daarin mee en staat dat er expliciet bij.

Bij een foute score is een fouttype toegekend. Verzonnen: het feit is nooit waar geweest; het product, de dienst, het bedrag of de voorwaarde bestaat niet en bestond niet. Verouderd: het feit was ooit waar maar klopt op de peildatum niet meer. Merkverwarring: het feit hoort bij een ander merk, het moederbedrijf of een keten met dezelfde naam. Overig: een fout die niet in de andere drie past. Verouderd en verzonnen worden overal als aparte categorieën gerapporteerd.

De kernbeslisregels bij het scoren: bij een ja- of nee-vraag telt de hele strekking van het antwoord, niet het eerste woord; bedragen en eenheden worden waar nodig omgerekend voordat ze worden vergeleken; extra details verlagen een score niet zolang ze niet misleidend zijn; en een slag om de arm of hedging is geen correctie van een onjuiste kern.

De scoring is ondersteund door een dubbele blinde judge-pipeline en afgesloten door een menselijke beoordelaar. Een eerste judge (gpt-5.4-nano) scoorde blind voor, zonder de modelnaam te kennen. Een tweede judge uit een andere modelfamilie (claude-haiku-4-5) scoorde met exact dezelfde prompt en regels, eveneens blind. Waar beide judges het eens waren en de uitkomst feitelijk juist was, is die consensus automatisch overgenomen; al het overige, plus alle strik- en controlerijen, ging handmatig door de review-queue. De kritieke poort was de vals-juist-cel: gevallen waar de menselijke beoordelaar fout of geen antwoord gaf terwijl een judge juist met hoge zekerheid gaf. Op de validatieset was die cel nul.

De overeenstemmingscijfers op de validatieset: judge 1 kwam op 80,4% overeen met de menselijke beoordelaar, judge 2 op 85,6% (feitelijk 84,2%, strik en controle 94,4%). Waar beide judges het eens waren, kwam de consensus op 92,1% overeen op de feitelijke rijen; consensus-juist en consensus-ja waren 100% zuiver op de validatieset. Een leniency-analyse zonder familievoorkeur liet zien dat de eerste judge antwoorden het minst omhoog corrigeert bij OpenAI: omhoog-correcties ChatGPT 1,3%, Gemini 0,0%, Perplexity 5,1%. Elke foutclaim en alle strik- en controlerijen zijn uiteindelijk door de menselijke beoordelaar bevestigd.

3.4 App-validatie

De kopcijfers komen uit de API. Om te toetsen of de chat-apps hetzelfde beeld geven, zijn 90 handmatige app-metingen gedaan (30 vragen maal 3 apps). Die kwamen voor 88,9% overeen met de API-meerderheid. De vooraf vastgelegde drempel van 90% is daarmee net niet gehaald. Van de tien afwijkingen zijn er vijf toe te schrijven aan gewone run-variatie, waarbij de API-runs onderling al niet eenduidig waren, en vijf aan een echt methodeverschil tussen app en API, dus vijf op negentig paren of 5,6%.

Als leescontext hoort daar een ruisplafond bij. Binnen de API stemt een losse run gemiddeld 96,3% overeen met de eigen model-meerderheid (ChatGPT 95,6%, Gemini 98,9%, Perplexity 94,4%). Dat is de theoretische bovengrens voor een enkele app-run tegen de API-meerderheid; 100% is door de eigen variabiliteit van de modellen niet haalbaar. De consequentie voor de claims: kopclaims worden geformuleerd als API-metingen in consumentenconfiguratie, en uitspraken over de apps krijgen deze validatie expliciet mee. Bij de vijf gedocumenteerde methodeverschillen is de app-observatie leidend; het benoemde voorbeeld is de Picnic-studentenpas, waar de Perplexity-app een valse premisse bevestigde die de API in drie runs corrigeerde.

3.5 Wijzigingslog

Tijdens het beoordelen bleken twaalf van de 185 geverifieerde feiten alweer achterhaald, hoewel ze enkele dagen eerder met de hand waren gecontroleerd. Ze zijn herzien na consistente signalen uit meerdere modellen en

hernieuwde verificatie tegen de primaire bron. In verschillende gevallen liepen de zoekende modellen daarmee voor op de handmatige verificatie. Twee andere grondwaarheden zijn na hercontrole juist bevestigd (NS en Jumbo). De tabel hieronder beschrijft de aard van elke wijziging; de exacte oude en nieuwe waarden vallen onder de niet-gepubliceerde grondwaarheden en staan in de aanvraagversie van de meetlog.

Vraag	Merk	Aard van de wijziging	Bron	Datum
O1-019	HEMA	Verzekeringen niet langer in eigen aanbod; overgedragen aan partners	eigen aanbodpagina	7 juli 2026
O1-043	KPN	Sim-only aanbod herzien (onbeperkt-abonnement met faire daglimiet)	kpn.com	7 juli 2026
O1-045	KPN	Hoogte combivoordeel herzien	kpn.com	7 juli 2026
O1-046	Vodafone	Maandtarief herzien	vodafone.nl	7 juli 2026
O1-062	ABN AMRO	Maandtarief zakelijk pakket herzien	eigen tariefpagina	7 juli 2026
O1-074	bunq	Spaarrentestructuur herzien (basis- en bonusrente)	eigen rentepagina	7 juli 2026
O1-081	KLM	Handbagageregels herzien (afhankelijk van tickettype)	klm.nl	7 juli 2026
O1-099	Bitvavo	Vergunningstatus herzien (MiCA via AFM in plaats van eerdere registratie)	AFM-register	7 juli 2026
O1-104	Knab	Juridische status herzien (bijkantoor onder buitenlandse bankvergunning)	officiele statusmelding	7 juli 2026
O1-142	Essent	Vaste leveringskosten herzien	eigen tariefpagina	7 juli 2026
O1-221	Moneybird	Pakketnamen en prijzen herzien	eigen prijspagina	7 juli 2026
O1-222	Moneybird	Gratis pakket vervallen; proefperiode in de plaats	eigen prijspagina	7 juli 2026

Twee grondwaarheden zijn na hercontrole bevestigd zonder wijziging: O1-076 (NS) en O1-034 (Jumbo).

4. Resultaten

Alle cijfers in dit hoofdstuk komen letterlijk uit het analyseblad van het werkboek of uit de herberekende bronnenanalyse. Het vraagniveau is het primaire cijfer; runniveau staat gelabeld waar het wordt gebruikt.

4.1 Kopcijfers per model (vraagniveau)

Meerderheidsoordeel van drie runs per vraag, basis 184 vragen per model. De betrouwbaarheidsintervallen zijn Wilson-intervallen van 95% op het foutpercentage.

Model	Juist	Deels	Fout	Geen antwoord	Niet volledig juist	Fout, 95%-BI
ChatGPT	94,6%	2,7%	2,7%	0,0%	5,4%	1,2 tot 6,2%
Gemini	94,0%	4,9%	1,1%	0,0%	6,0%	0,3 tot 3,9%
Perplexity	83,7%	6,0%	9,2%	1,1%	16,3%	5,8 tot 14,3%

Niet volledig juist telt deels, fout en geen antwoord samen; deels is geen fout. De intervallen van Perplexity en Gemini overlappen niet, wat de spreiding van ongeveer een factor 8 tussen het slechtste en het beste model onderbouwt. Ter sensitiviteit op runniveau (runs binnen een vraag zijn niet onafhankelijk) liggen de foutpercentages op 3,1%, 1,4% en 10,3%.

4.2 Aard van de fouten per model (runniveau)

Fouttypes liggen van nature op runniveau, omdat een type per foute run wordt toegekend. Onderstaande verdeling telt alleen runs met de score fout.

Model	Fout-runs	Verouderd	Verzonnen	Merkverwarring	Overig
ChatGPT	17	10	0	0	7
Gemini	8	7	0	0	1
Perplexity	57	34	9	1	13
Totaal	82	51	9	1	21

Over alle foute runs is 62% verouderd (51 van 82) en 11% verzonnen (9 van 82). Verzonnen fouten komen uitsluitend bij Perplexity voor. Verouderd is bij elk model het grootste fouttype.

4.3 Tijdgevoelige versus stabiele feiten

Op vraagniveau, meerderheidsoordeel, geteld over vraag-modelcombinaties (79 tijdgevoelige en 106 stabiele vragen, elk maal drie modellen).

Type feit	Fout-meerderheid	Aandeel verouderd binnen de foute meerderheidsvragen
Tijdgevoelig	6,8% (16 van 237)	88% (14 van 16)
Stabiel	2,5% (8 van 318)	25% (2 van 8)

Tijdgevoelige feiten gaan ruim tweeënhalf keer zo vaak fout, en waar het misgaat is het bijna altijd verouderde informatie. Gemini maakte nul fouten op de 106 stabiele vragen. Dit bevestigt H4.

4.4 Consistentie

Aandeel vragen waarbij de drie identieke runs hetzelfde oordeel opleverden, per model (vraagniveau).

Model	Consistente vragen
ChatGPT	91,9% (170 van 185)
Gemini	91,4% (169 van 185)
Perplexity	78,9% (146 van 185)

Consistentie staat los van juistheid: een model kan consistent hetzelfde foute antwoord geven. Perplexity is niet alleen vaker fout, maar ook wisselvalliger tussen runs.

4.5 Strikvragen en controles

Meerderheidsoordeel per model, 14 strikvragen en 14 controles per model. Bij een strikvraag is ja slecht (de valse premisse wordt bevestigd); bij een controle is ja goed (de ware premisse wordt terecht bevestigd).

Model	Strik: valse premisse bevestigd	Strik 95%-BI	Controle: ware premisse terecht bevestigd
ChatGPT	7,1% (1 van 14)	1,3 tot 31,5%	100% (14 van 14)
Gemini	0,0% (0 van 14)	0,0 tot 21,5%	100% (14 van 14)
Perplexity	7,1% (1 van 14)	1,3 tot 31,5%	92,9% (13 van 14)

Bevestigd betekent hier een meerderheid ja of deels. De enige ingetrapte ChatGPT-strikvraag had een deels-meerderheid (een run ja, twee runs deels) en telt daarmee als bevestigd; op vraagniveau bevestigen ChatGPT en Perplexity elk een van de veertien valse premisses (samen 2 van 42 vraag-modelcombinaties, zie 4.7).

Ter sensitiviteit op runniveau (42 strik- en 42 controlemetingen per model): ChatGPT trapte in 3 van 42 strikmetingen (7,1%, interval 2,5 tot 19,0%) en had 42 van 42 controles goed; Gemini 0 van 42 strik (interval 0,0 tot 8,4%) en 42 van 42 controles; Perplexity 5 van 42 strik (11,9%, interval 5,2 tot 25,0%) en 38 van 42 controles goed. De aantallen zijn klein en de intervallen overlappen, waardoor H5 onbeslist blijft.

4.6 Bronnen die in de antwoorden worden aangehaald

De bronanalyse ligt van nature op runniveau. De basis is per antwoord de verzameling unieke domeinen die het model als bron toonde. Voor Gemini staan de echte domeinen in de titel van de bronverwijzing, omdat de API-url een Google-redirect is; die zijn zo genormaliseerd. Domeinen zijn genormaliseerd naar hun hoofddomein en per merk geclassificeerd. Formulering overal: het gaat om bronnen die in foute antwoorden worden aangehaald, niet om een bewezen oorzaak van de fout.

Dekking: ChatGPT toonde bij 550 van 555 antwoorden minstens een bron (99,1%), Gemini bij 554 van 555 (99,8%) en Perplexity bij 555 van 555 (100%).

Aanwezigheid van het eigen merkdomein (staat het merkdomein tussen de bronnen), juiste tegen foute antwoorden:

Model	Juiste antwoorden	Foute antwoorden
ChatGPT	94,3% (494 van 524)	87,1% (27 van 31)
Gemini	98,8% (510 van 516)	97,4% (37 van 38)
Perplexity	96,7% (437 van 452)	90,8% (89 van 98)

Aanwezigheid en aandeel zijn niet hetzelfde. Aanwezigheid is of het merkdomein er tussen staat; aandeel is hoeveel van alle aangehaalde bronnen van dat merkdomein komen. De categorieverdeling hieronder toont het aandeel van de citaties per categorie, per model, voor juiste en foute antwoorden.

Categorie	ChatGPT juist	ChatGPT fout	Gemini juist	Gemini fout	Perplexity juist	Perplexity fout
eigen merkdomein	81,8	55,1	25,5	15,3	18,9	16,4
buitenlandse merkvariant	1,6	8,2	2,0	0,4	1,6	0,0
vergelijker of affiliate	4,1	20,4	23,0	39,8	29,2	30,6
nieuws of media	1,3	2,0	12,7	8,0	12,6	13,3
social of fora	0,0	0,0	2,6	1,5	11,9	12,3
overheid of toezichthouder	3,1	2,0	0,7	0,4	1,3	0,2
overig	8,0	12,2	33,5	34,7	24,5	27,3

Alle waarden zijn percentages van de bron-citaties in die groep. Aantal citaties: ChatGPT juist 611 en fout 49, Gemini juist 2237 en fout 274, Perplexity juist 2646 en fout 579. Bij ChatGPT en Gemini verschuift het bronprofiel in foute antwoorden sterk richting vergelijkers en affiliatesites (van 4,1 naar 20,4% en van 23,0 naar 39,8%); Perplexity leunt in beide gevallen het zwaarst op bronnen van derden, met social en fora als eigen blok van ongeveer 12%.

De meest aangehaalde derde-partij-domeinen in foute antwoorden, per model:

Model	Meest aangehaalde derden in foute antwoorden
ChatGPT	klm.com (3), transportation.gov (2), gaslicht.com (2), energievergelijk.nl (2)
Gemini	overstappen.nl (11), independer.nl (10), easyswitch.nl (6), mobiel.nl (5), tweakers.net (5), keuze.nl (5), energievergelijk.nl (5), energiekiezer.nl (5)
Perplexity	reddit.com (18), facebook.com (17), instagram.com (12), belsimpel.nl (11), mobiel.nl (9), seniorweb.nl (8), overstappen.nl (8), pricewise.nl (8)

Bij ChatGPT is het aantal bron-citaties in foute antwoorden klein (49); getoond zijn de derde-partij-domeinen met minstens twee foute runs. Classificatieverantwoording: de 464 domeinen die in drie of meer antwoorden voorkomen (samen 90,6% van alle citaties) zijn handmatig ingedeeld; domeinen met minder dan drie voorkomens vallen onder overig (afkapregel). De volledige domeinclassificatie is als los bestand beschikbaar.

4.7 Hypothese-uitkomsten

Vraagniveau primair; runniveau alleen als gelabelde sensitiviteit. De formuleringen zijn de letterlijke hypotheses uit het preregistratieblad.

Hypothese	Uitkomst	Onderbouwing (vraagniveau)
H1: modellen geven op een substantieel deel een onjuist antwoord	Verworpen in de algemene vorm	Fout 2,7 / 1,1 / 9,2%; alleen Perplexity substantieel
H2: kleinere merken leveren meer foute antwoorden op	Verworpen	Geen patroon groot naar mkb bij enig model; omkering bij Perplexity niet significant (z is 1,05, p is 0,30)
H3: geen antwoord komt vaker voor bij mkb-merken	Niet ondersteund	Geen-antwoord-meerderheid bij 2 van 555 vraag-modelcombinaties
H4: tijdgevoelige feiten gaan vaker fout	Bevestigd	Tijdgevoelig 6,8% tegen stabiel 2,5%; verouderd 88 tegen 25%
H5: valse premisses vaker bevestigd dan ware ontkend	Onbeslist	Strik 2 van 42 tegen controle 1 van 42 op vraagniveau; kleine n, intervallen overlappen
H6 (exploratief): antwoorden zonder bron zijn vaker fout	Niet toetsbaar	1659 van 1665 runs toonden een bron; er is geen zonder-bron-groep

5. Acht cases

Dit hoofdstuk is de enige plek met letterlijke antwoordfragmenten. De fragmenten zijn ingekort tot de relevante passage. De vragen zijn geparafraseerd; de exacte vraagteksten en grondwaarheden worden niet gepubliceerd. Elke case volgt hetzelfde stramien: vraag, antwoordfragment, aangehaalde bronnen, oordeel en fouttype, les.

Case 1. Rechtstreex: een gestopt bedrijf leeft online door

Vraag (geparafraseerd): een strikvraag over de bezorg- of servicekosten bij Rechtstreex, een lokale boodschappendienst.

Antwoordfragment (Perplexity): "Bij Rechtstreex betaal je geen aparte bezorg- of servicekosten; deze kosten zijn inbegrepen in de eindprijs van de producten."

Aangehaalde bronnen: rechtstreex.nl, stapjebeter.nl, slowfood.nl, youtube.com, facebook.com.

Oordeel en fouttype: fout, verouderd. Rechtstreex is sinds oktober 2025 failliet en levert niet meer; Perplexity beantwoordde de kostenvraag in alle drie de runs alsof het bedrijf gewoon actief was. ChatGPT en Gemini gaven correct aan dat het bedrijf is gestopt.

Les: een gestopt bedrijf of product leeft online door tot iemand de bronnen opruimt. Zolang oude pagina's blijven staan, blijven ze in de antwoorden aangehaald worden.

Case 2. Picnic-studentenpas: app tegen API

Vraag (geparafraseerd): een strikvraag over een niet-bestaande Picnic-studentenpas met studentenkorting.

Antwoordfragment (ChatGPT, bevestigende run): "De Picnic-studentenpas geeft 20% korting op je boodschappen, dit wordt vermeld als studentenkorting via UNiDAYS."

Aangehaalde bronnen: onder meer kortingssites zoals diski.nl; de officiele Picnic-bron ontkent de korting.

Oordeel en fouttype: strikvraag, valse premisse deels of geheel bevestigd in enkele runs. Belangrijker is het app-versus-API-verschil: de Perplexity-app bevestigde de korting, terwijl de Perplexity-API die in drie runs juist ontkende. Dit is een van de vijf gedocumenteerde methodeverschillen, waarbij de app-observatie leidend is.

Les: affiliate- en kortingscontent bepaalt mede wat AI over je merk zegt, en de app en de API van hetzelfde model kunnen verschillen.

Case 3. Winkel 43: het risico is tweezijdig

Vraag (geparafraseerd): een strikvraag over het reserveren van een tafel bij Winkel 43, met een gespiegelde controle over het vooraf bestellen van een hele appeltaart.

Antwoordfragment (Perplexity, strik, foute run): "Je kunt een tafel reserveren bij Winkel 43 via hun officiele website." Antwoordfragment (Perplexity, controle, foute run): "Nee, het klopt niet dat je bij Winkel 43 een hele appeltaart vooraf kunt bestellen."

Aangehaalde bronnen: winkel43.nl naast reisgidsen en social zoals tripadvisor, instagram en reddit.

Oordeel en fouttype: bij Winkel 43 kan geen tafel worden gereserveerd, maar wel een taart worden besteld. Perplexity bevestigde in een run de niet-bestaande reservering en ontkende in meerdere runs de wel bestaande taartbestelling. ChatGPT en Gemini hadden beide kanten juist.

Les: het risico is tweezijdig. Een model kan niet alleen iets verzinnen, maar ook bestaand aanbod ten onrechte ontkennen.

Case 4. MediaMarkt: derden vertellen het verhaal

Vraag (geparafraseerd): feitelijke vragen over garantie, retour en bezorging bij MediaMarkt.

Aangehaalde bronnen in de foute Perplexity-antwoorden: radar-forum.avrotros.nl, reddit.com, instagram.com, test-aankoop.be, consumentenbond.nl en gathering.tweakers.net; mediamarkt.nl zelf werd in deze foute antwoorden vrijwel niet aangehaald.

Oordeel en fouttype: fout, verouderd, en in een geval verwarring met een gelijknamige buitenlandse partij. De feitelijke informatie staat gewoon op de eigen site van MediaMarkt.

Les: wie zijn eigen feiten niet goed vindbaar publiceert, laat derden het verhaal vertellen. Ook een groot merk verliest dan de regie over zijn eigen feiten.

Case 5. KLM: een Amerikaanse regel op een Nederlandse vraag

Vraag (geparafraseerd): kan een KLM-ticket binnen 24 uur na boeking kosteloos worden geannuleerd.

Antwoordfragment (ChatGPT): "Voor vluchten van of naar de VS: ja, KLM vermeldt dat je alleen voor vluchten naar of vanuit de VS, als je via klm.us hebt geboekt, binnen 24 uur na boeking een terugbetaling kunt aanvragen."

Aangehaalde bronnen: klm.com (de Amerikaanse KLM-site) in alle drie de runs, en transportation.gov, de site van het Amerikaanse ministerie van Transport, in twee van de drie.

Oordeel en fouttype: fout, overig. ChatGPT gaf in alle drie de runs een fout antwoord door de Amerikaanse 24-uursregel, die alleen geldt voor vluchten naar of vanuit de Verenigde Staten, te presenteren als antwoord op een Nederlandse vraag.

Les: modellen kunnen een buitenlandse regel importeren die niet op de Nederlandse situatie van toepassing is. Het gaat om bronnen die bij het foute antwoord staan, niet om een aangetoonde oorzaak.

Case 6. Moneybird: verouderd binnen dagen

Vraag (geparafraseerd): heeft Moneybird een gratis pakket, en wat kosten de pakketten.

Antwoordfragment (Gemini, foute run): "Ja, Moneybird heeft een gratis versie (het gratis pakket)."

Aangehaalde bronnen: de eigen Moneybird-domeinen en enkele boekhoudvergelijkers.

Oordeel en fouttype: fout, verouderd. Moneybird schrapte het gratis pakket rond 1 juli 2026 en herzag de pakketnamen en prijzen. De meeste runs van de zoekende modellen hadden de wijziging al opgepikt en waren juist; enkele runs gaven nog het oude gratis pakket. Dit is een van de twaalf grondwaarheden die tijdens de meetweek zijn herzien.

Les: tijdgevoelige feiten kunnen binnen dagen verouderen. Een merk dat een wijziging niet duidelijk op de eigen site zet, riskeert dat de oude versie blijft rondzingen.

Case 7. HEMA: een gestopt aanbod dat blijft hangen

Vraag (geparafraseerd): verkoopt HEMA verzekeringen.

Antwoordfragment (ChatGPT, foute run): "Ja, HEMA heeft verzekeringen onder HEMA Verzekeringen."
Antwoordfragment (Gemini, juiste run): "Nee, HEMA verkoopt op dit moment geen verzekeringen of financiële producten meer."

Aangehaalde bronnen: een mix van eigen en derde-partij-pagina's.

Oordeel en fouttype: fout of deels, verouderd. HEMA heeft het verzekeringsaanbod afgestoten; de schadeverzekeringen zijn overgedragen en de zorgverzekering is bij een andere partij ondergebracht. ChatGPT bevestigde in een run nog het oude aanbod, terwijl Gemini de actuele situatie juist had.

Les: een gestopt aanbod blijft in AI-antwoorden hangen zolang oude bronnen dat aanbod nog noemen. Meld gestopte producten expliciet af op je eigen site.

Case 8. Action: het enige verzonnen patroon

Vraag (geparafraseerd): kun je bij Action online reserveren en ophalen (click en collect), en hoe kun je er betalen.

Antwoordfragment (Perplexity): "Ja, je kunt bij Action in Nederland online producten reserveren via Click en Collect om ze vervolgens in de winkel op te halen." En in een andere vraag: "Action weigert wettig Nederlands betaalmiddel in contanten."

Aangehaalde bronnen: een mix van retailnieuws, social en een verkeerd gekoppeld buitenlands domein (actionpay.id).

Oordeel en fouttype: fout, verzonnen. Beide beweringen zijn nooit waar geweest. Dit past in het enige verzonnen patroon in de meting: alle negen verzonnen fouten komen van Perplexity, en bij alle negen stond het eigen merkdomein van het merk gewoon tussen de aangehaalde bronnen.

Les: verzonnen fouten zijn zeldzaam en modelspecifiek. Dat het juiste merkdomein wordt aangehaald, is geen garantie dat het antwoord klopt.

6. Limitaties

Dit onderzoek is een momentopname binnen een week. De metingen zijn uitgevoerd op 5 en 6 juli 2026, met grondwaarheden geverifieerd op peildatum 4 juli 2026 en waar nodig herzien op 7 juli 2026; feiten die daarna veranderen, vallen buiten de meting. De meetweek zelf liet al zien hoe snel dat kan gaan, met twaalf herziene grondwaarheden.

De meting gebruikt de API in de standaard consumentenconfiguratie, zonder deep research of expliciete redeneerstanden. De resultaten gelden voor die configuratie. De app-validatie kwam op 88,9% overeen met de API, net onder de vooraf gestelde drempel van 90%, waarvan ongeveer de helft van de afwijkingen gewone run-variatie was. App en API kunnen dus verschillen.

Per merk zijn vijf vragen gesteld. Dat is genoeg voor een uitspraak over modellen en over vraagcategorieën, maar te weinig voor een ranglijst van merken; de per-merk-cijfers zijn indicatief en geen oordeel over een individueel merk. De tiers verschillen bovendien in vraagsamenstelling, wat een zuivere tiervergelijking bemoeilijkt.

De categorie deels is heterogeen: ze bundelt antwoorden met een juiste kern maar een onvolledig of licht onjuist element. Deze categorie telt niet als fout. Waar een niet-volledig-juist-cijfer wordt genoemd, staat erbij dat deels daarin meetelt.

De bronvermelding is geen causaliteit. Dat een bepaald type bron in foute antwoorden vaker wordt aangehaald, bewijst niet dat die bron de fout heeft veroorzaakt; we rapporteren consequent als bronnen die in foute antwoorden worden aangehaald.

De scoring is ondersteund door judges, maar de menselijke beoordelaar heeft de eindverantwoordelijkheid; elke foutclaim en alle strik- en controlerijen zijn met de hand bevestigd. H6, over antwoorden zonder getoonde bron, is niet toetsbaar omdat vrijwel alle antwoorden een bron toonden.

7. Vooruitblik

Twee vervolgen staan op de rol. Het eerste is O1b, een no-search-meting: dezelfde vragen zonder de zoekfunctie van de modellen, om te scheiden wat uit training komt en wat uit de live webindex. Die meting is met 1278 runs al uitgevoerd; publicatie volgt. Het tweede is een herhaalmeting met een deels vernieuwde vragenset, omdat gepubliceerde vragen en cases als verbrand gelden zodra ze openbaar zijn en modellen erop getraind kunnen raken.

8. Verantwoording en beschikbaarheid

De data zijn beschikbaar in drie lagen. Laag 1 is dit document: de volledige methode, de geaggregeerde resultaten en de verantwoording. Laag 2 is het open databestand open_data_O1.csv, met per feitelijke meting de vraagcode, het model, de run, de vraagcategorieën, de score, het fouttype en of er een bron werd getoond, plus vlaggen voor de aanwezige broncategorieën; zonder vraagteksten, grondwaarheden of antwoordteksten. Bij dat bestand hoort een reproductiescript dat de kopcijfers van hoofdstuk 4 uit de open data herrekent. Laag 3 is de volledige meetlog inclusief vraagteksten en grondwaarheden; die staat niet online, maar is voor verificatiedoeleinden op aanvraag beschikbaar via info@timmermansmedia.nl.

De onderzoekspagina met de interactieve visuals staat op timmermansmedia.nl/onderzoek/ai-hallucinaties-nederlandse-merken/. De cijfers in dit document en op die pagina zijn dezelfde.